LETTERKENNY INSTITUTE OF TECHNOLOGY
ASSIGNMENT COVER SHEET

Lecturer's Name: Angela Sweeney

Assessment Title: Extract Transform Load(ETL) CA Lab Report

Work to be submitted to: Angela Sweeney

Date for submission of work: November 29, 2023

Place and time for submitting work: _____

To be completed by the Student

Student's Name: Arya Sasi

Class: MSc Big Data Analytics

Subject/Module: Business intelligence

Word Count (where applicable): _____

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: ARYA SASI Date: November 29, 2023

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero.

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment shall normally be carried forward from the original examination to the repeat examination.

Declaration:

I declare that this work is entirely my own and does not contain the words or ideas of someone else, whether published or not, without specific acknowledgement by relevant referencing. I have read and understood the LYIT Plagiarism Policy on the "Student & Academic Policies" section of the LYIT Website and understand plagiarism to include:

- Direct copying of text, images and other materials (electronic or otherwise) from a book, article, fellow student's essay, handout, web page or other source without proper acknowledgement.
- Claiming individual ideas derived from a book, article etc. as one's own and incorporating them into one's work without acknowledging the source of these ideas.
- Overly depending on the work of one or more other sources without proper acknowledgement of the source, by constructing an essay, project etc., extracting large sections of text from another source and merely linking these together with a few of one's own sentences.

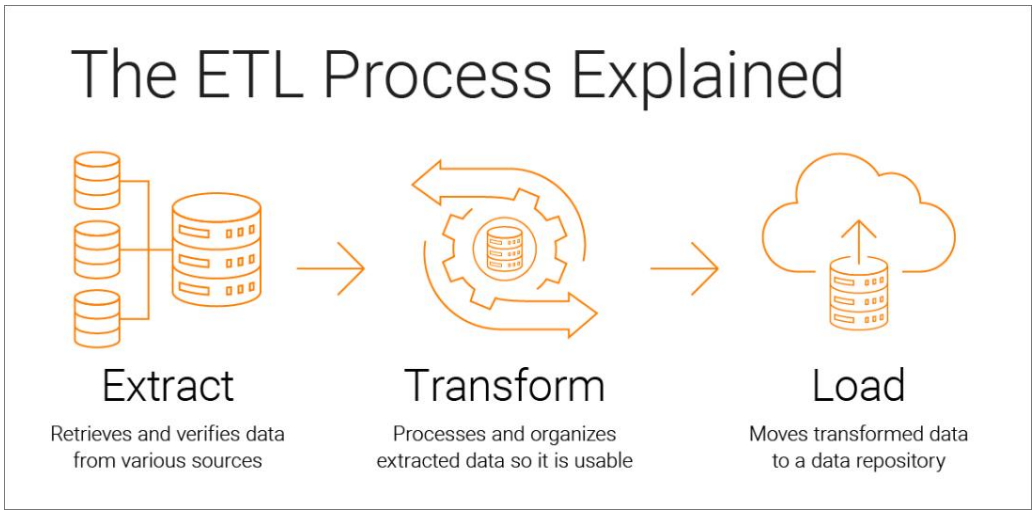
I understand that it is my responsibility to familiarise myself with and to follow the Institute's Assessment Regulations. I acknowledge that Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations and that penalties will be applied if I breach this policy.

Signed: ARYA SASI Date: November 29, 2023

Extract Transform Load

(ETL)

CA Lab Report



Description

ETL is the process of merging data from numerous sources into a big, centralised repository known as a data warehouse. ETL cleans and organises raw data in order to prepare it for storage, data analytics, and machine learning (ML).

A Slowly Changing Dimension (SCD) is a dimension that stores and manages both current and historical data over time in a data warehouse. It is considered and implemented as one of the most critical ETL tasks in tracking the history of dimension records.

This practical describes a Job that stores and manages both the current and historical Customer data in a MySQL table using SCD (Slowly Changing Dimensions).The input data contains various Customer details including their customer ID, customer first name ,customer last name,customer address ,pincode ,customer DOB and so on.Obtain the files called customer _details_ i.xls ,customer_ details_ iu.xls and customer_details_u.xls from the Practicals folder already given.

Objectives

This CA is intended to assess you on the following Learning outcomes

1. Create Mysql database for staging and dimensions tables.

2. Create Talend metadata for source, staging and dimension objects.

3. Cleanse the data to remove anomalies while loading from file to staging.

4. Implement SCD logic.

TASK 1

Objectives

Create Mysql database for staging and dimensions tables.

Method

Create MySql Schema called scd_test1 that contains a table called customer_detail_dim. Created and forwarded engineer the model as illustrated below (Figure 1) or use the script called scd_test1_schema.sql file.

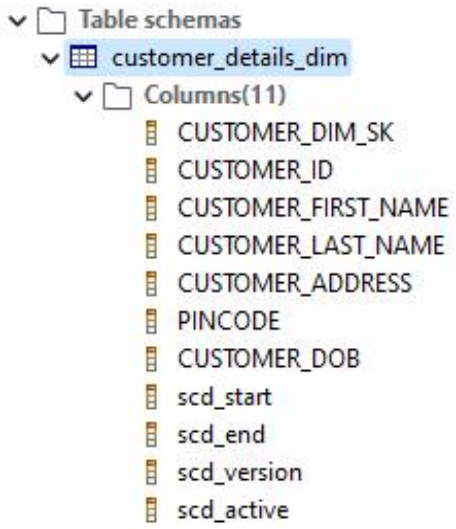


Figure 1 : Customer Details Attribute

For creating the schema.we need to open MySQL workbench and open a new sql file and name it as scd_test1_schema and script it as shown in Figure 2.

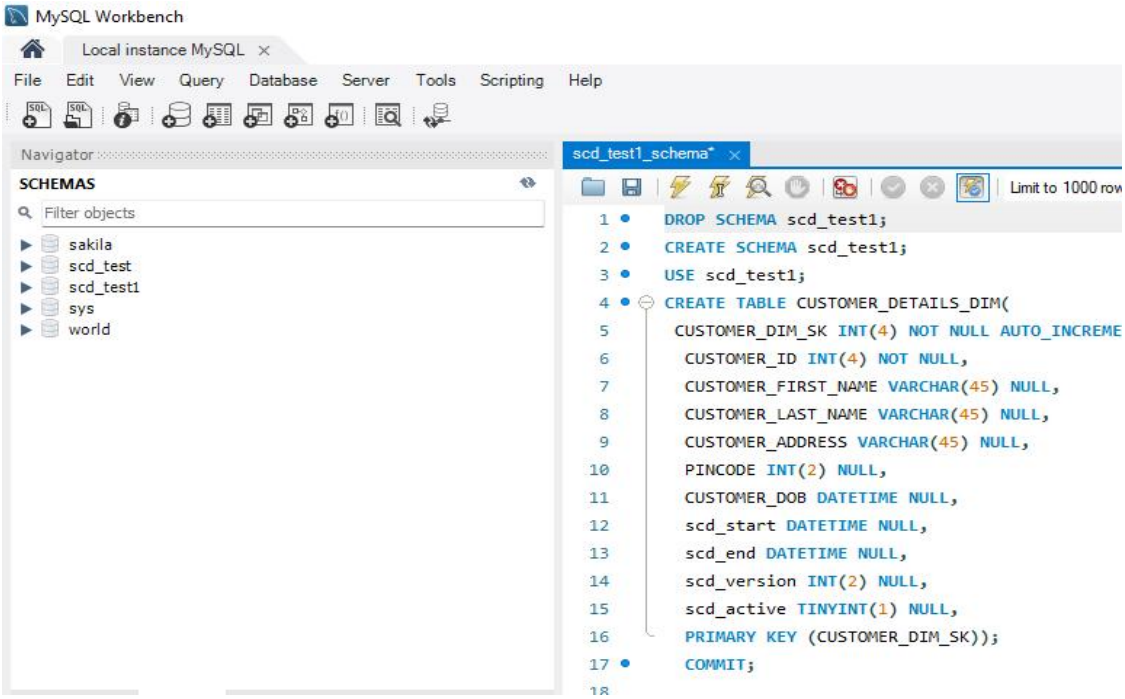


Figure 2 : CUSTOMER_DETAIL_DIM table creation.

For Storing the data cleaned we need to create a staging table.for that we need to create the table as shown below.

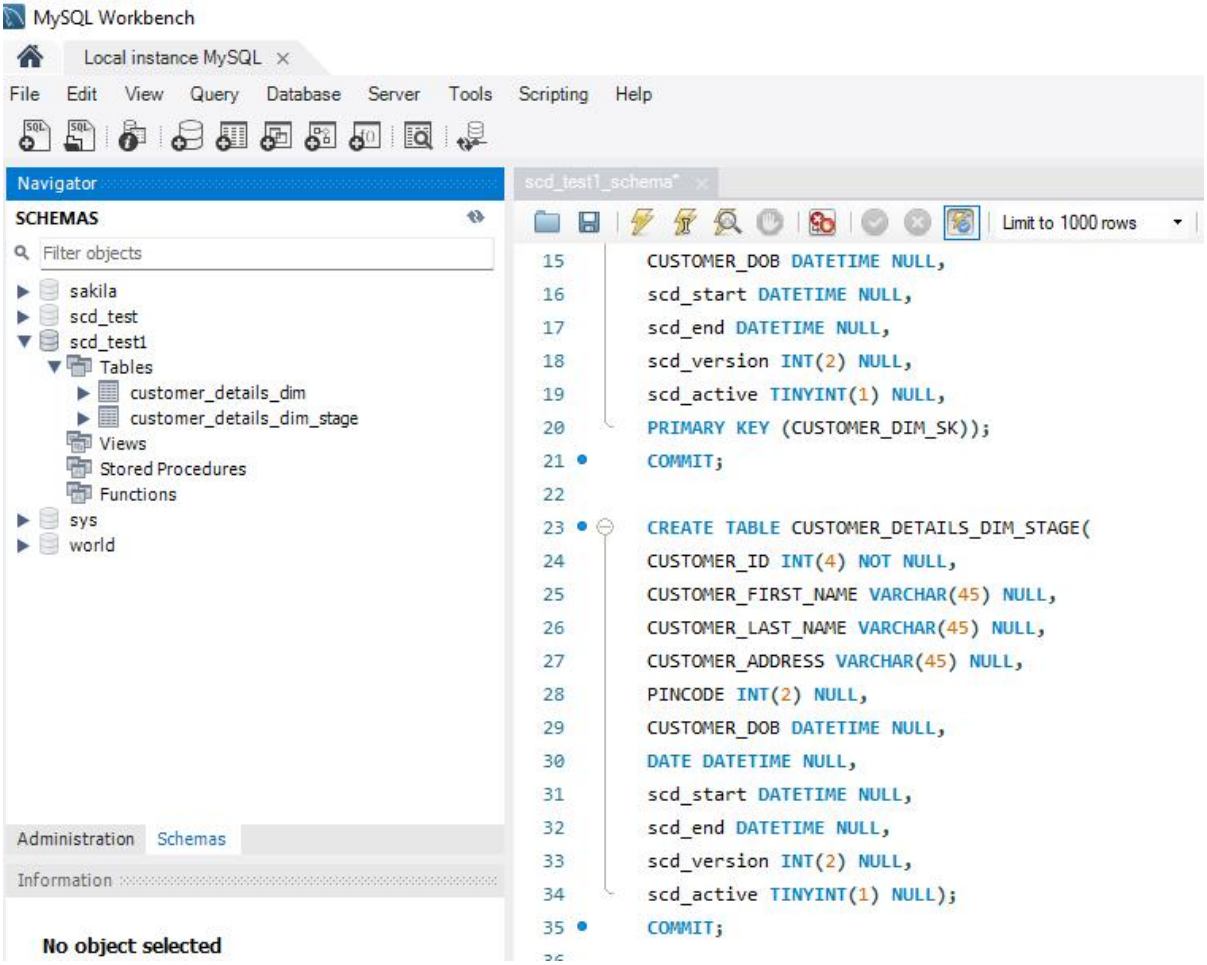


Figure 2.1 : Staging Table.

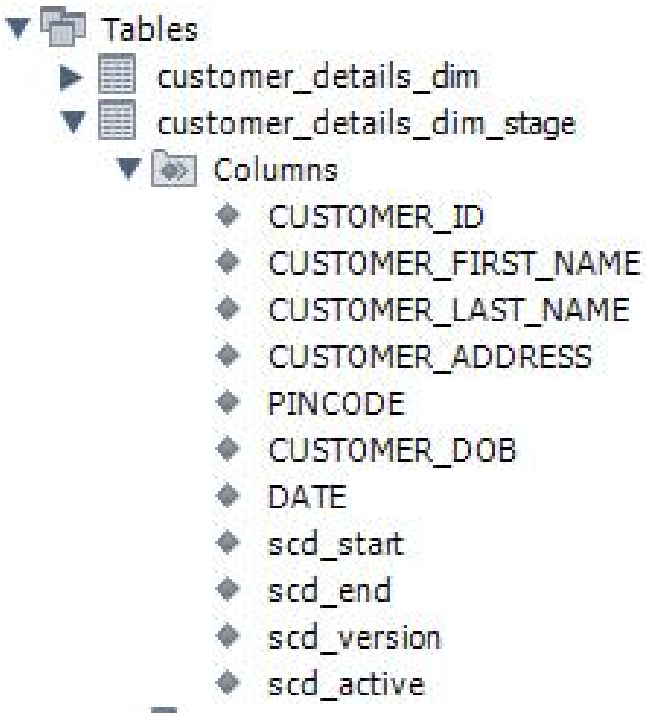


Figure 2.2 : Fields in staging Table.

Result

The schema : scd_test1 and the table : CUSTOMER_DETAILS_DIM are created in MySQL workbench as shown in below Figures

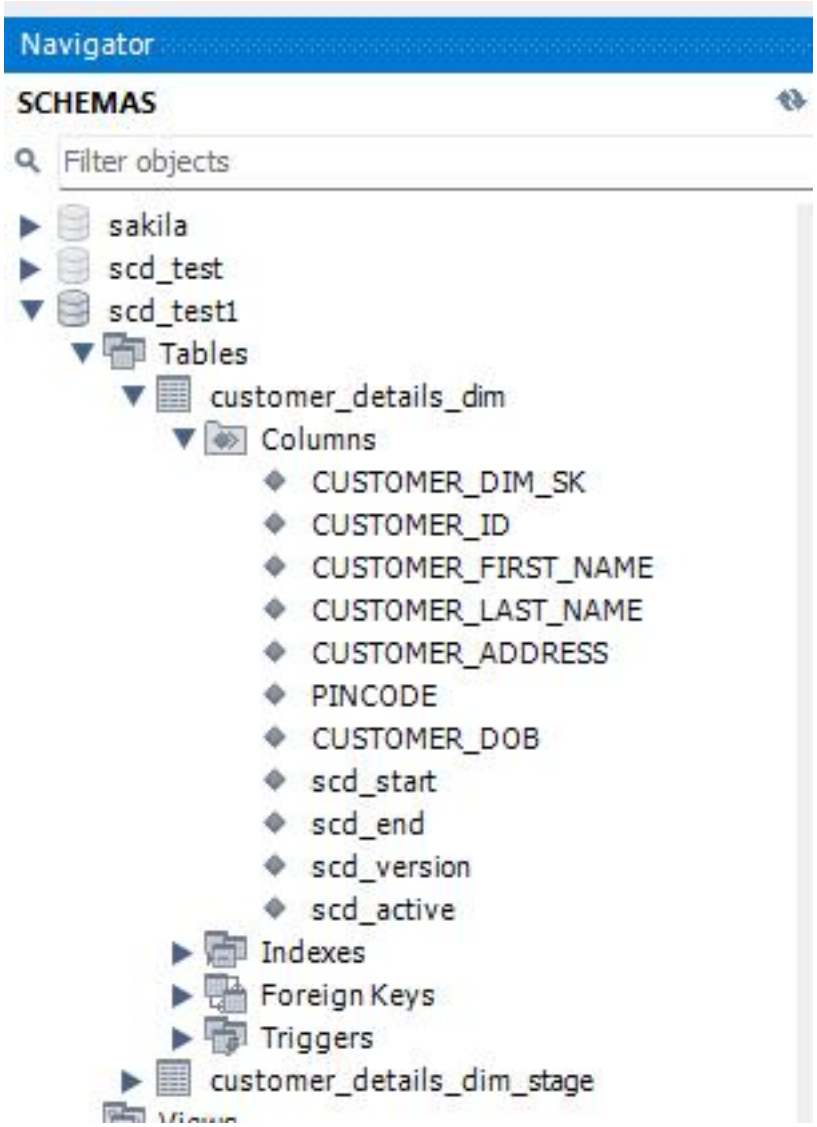


Figure 3 : created customer_details_dim schema and table.

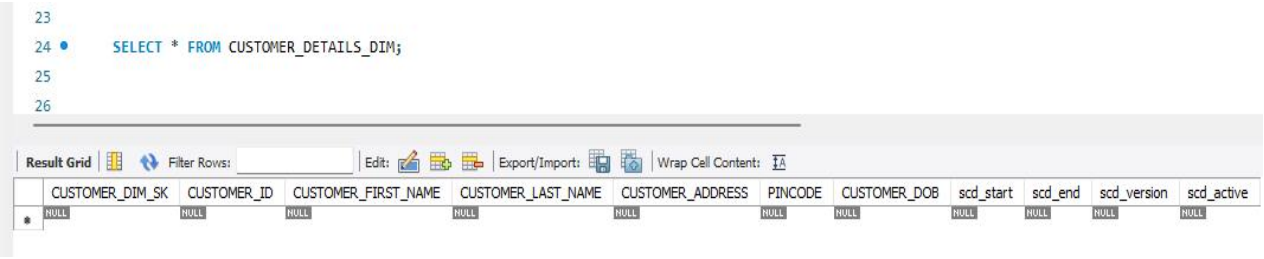


Figure 3.1 : CUSTOMER_DETAILS_DIM table view.

The schema : scd_test1 and the table : CUSTOMER_DETAILS_DIM_STAGE are created in MySQL workbench as shown in below Figures.

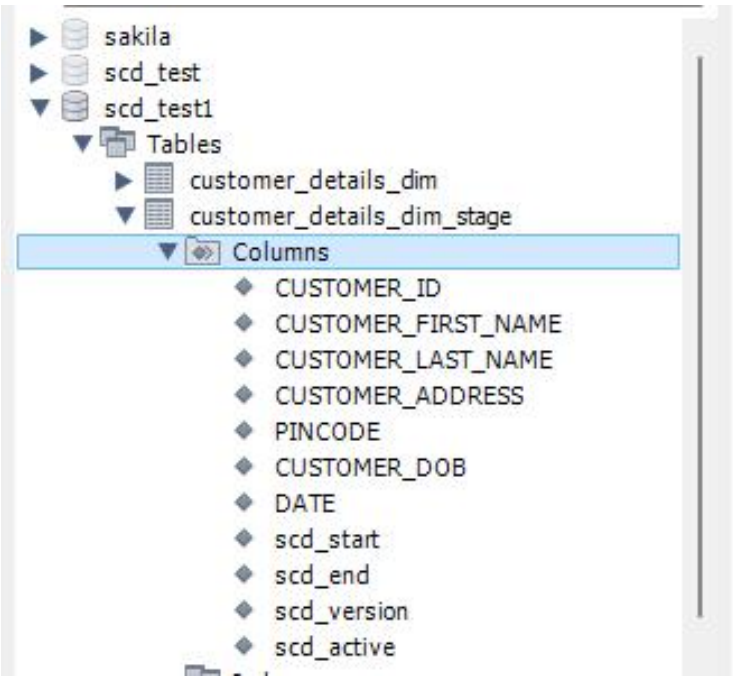


Figure 3.2 : created customer_details_dim_stage schema and table.

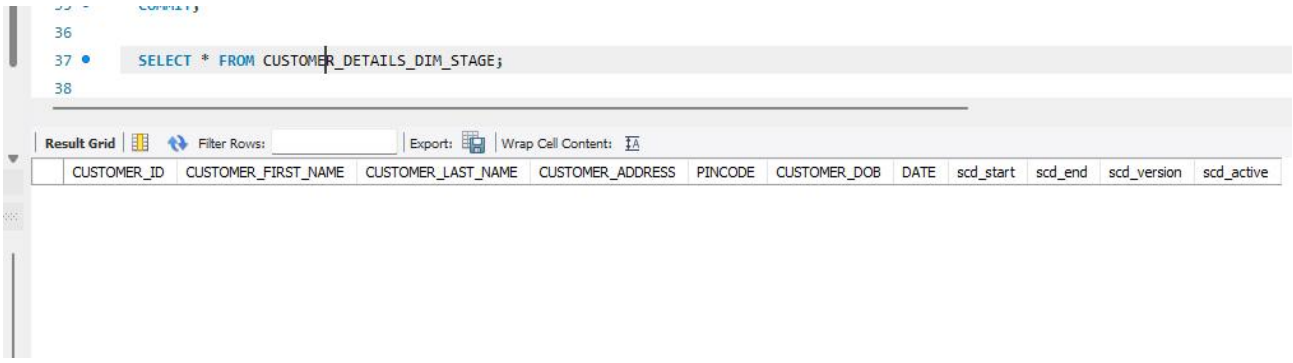


Figure 3.3 :CUSTOMER_DETAILS_DIM_STAGE table view.

TASK 2

Objectives

Create talend metadata for source, staging and dimension objects.

Method

The customer data has been extracted in.xlsx format. The files are as follows.

CUSTOMER_DETAIL_U	25/11/2023 15:49	XLSX File	9 KB
CUSTOMER_DETAIL_IU	25/11/2023 13:00	XLSX File	9 KB
CUSTOMER_DETAIL_I	25/11/2023 13:00	XLSX File	10 KB

Open talend open studio.Talend Open Studio(TOS) requires all jobs to be part of a project.So open a project.

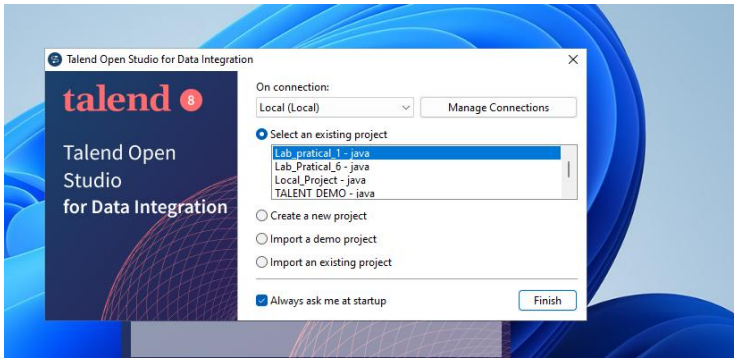


Figure 4 : Open talend open studio.

Then go to metadata on the sidebar and right click on the File Excel to create Excel file as show in Figure 4.1.

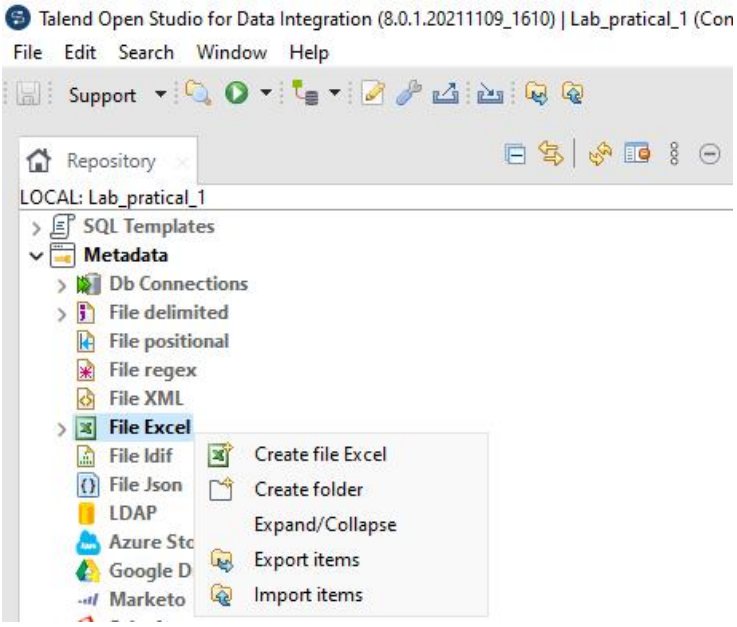


Figure 4.1 : Creating Metadata

After that do the following steps to create the metadata.

Step 1: Give a Name for the File and click on Next Button.

Figure 4.2 : Naming the file.

Step 2 : Import the excel file CUSTOMER_DETAILS_I.xlsx on clicking on browse option and do the same as shown below.Then click on next button.

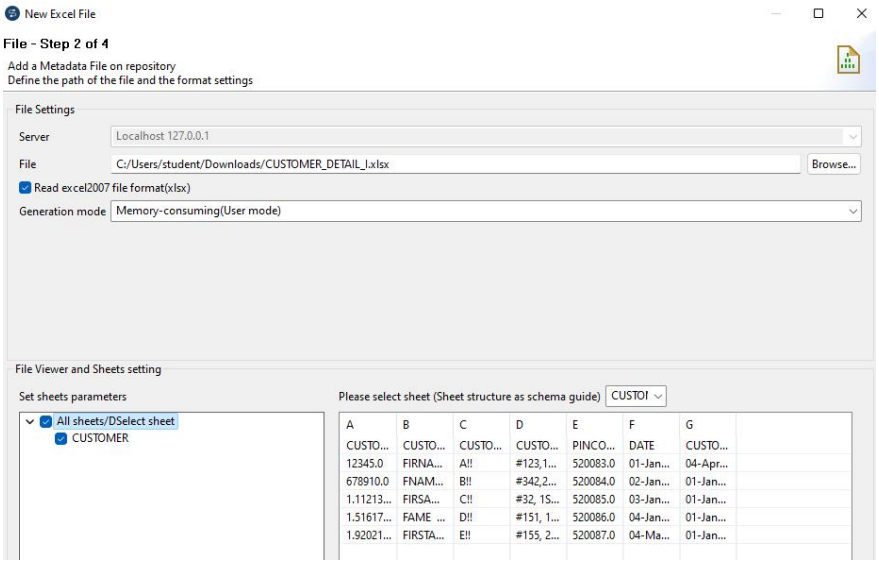


Figure 4.3 :Defining the path and format.

Step 3 :Then Check the data is same as shown in figure 4.4 and click on Next button.

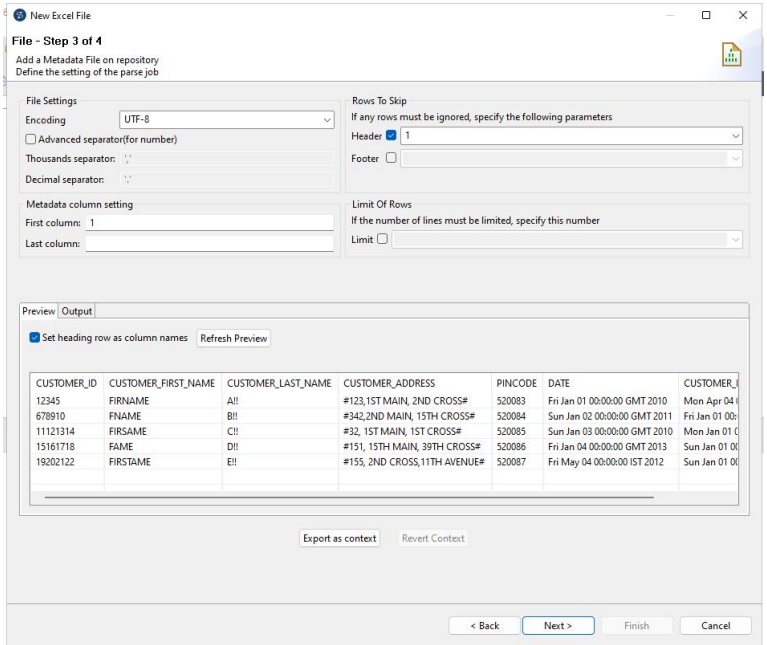


Figure 4.4 : Define the Settings

Step 4 : Define the Schema as follows and click on Finish Button.

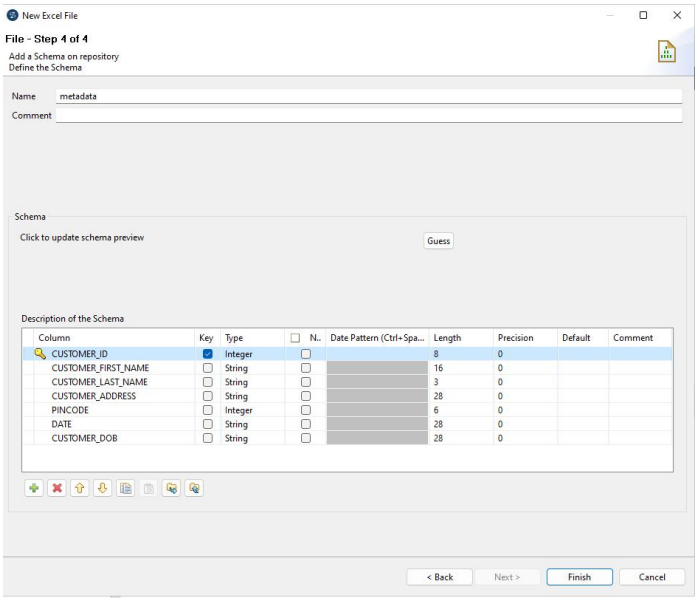
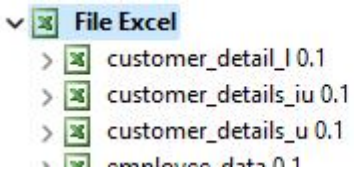


Figure 4.5 : Define Schema.

After this the Metadata for the Excel File CUSTOMER_DETAILS_I.xlsx is created.Do the Same for CUSTOMER_DETAILS_IU.xlsx and CUSTOMER_DETAILS_U.xlsx.

We should have Metadata as follows in your repository:



Create a DB Connection and Customer table metadata in Talend Open Studio.For that follow the instructions below to complete this task.

1. In the repository, select metadata and DB Connections

2. Right click on DB Connections to create a new connection. Enter SCD_TEST_CONNECT1 and other inputs (i.e. purpose and description) and click Next.

Figure 5 :Create a new database connection.

3. Complete the parameters in fig 6 below and test the connection. The password for the MySql server in the labs in password.

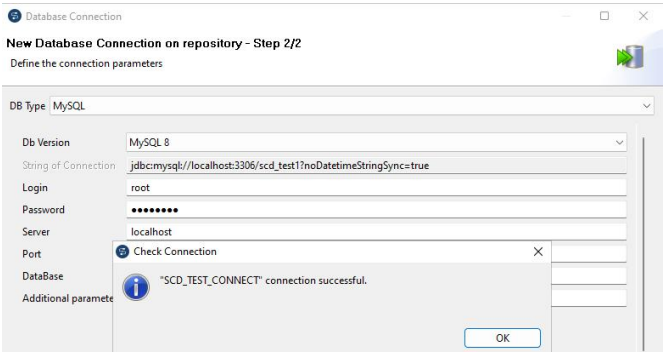


Figure 5.1 : Define the connection parameters

4. If the connection is successful then select finish to create it.

5. Now right click on the connection to retrieve the schema for the tables and click next.

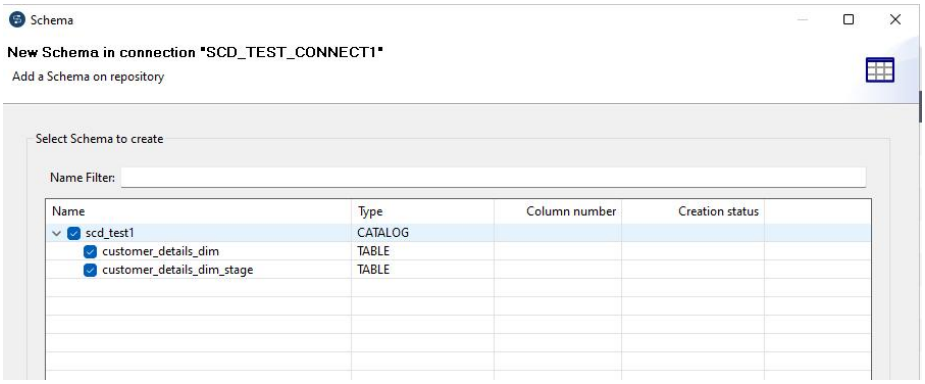


Figure 5.2 : Select tables in the database.

6. Check the columns and their datatypes.

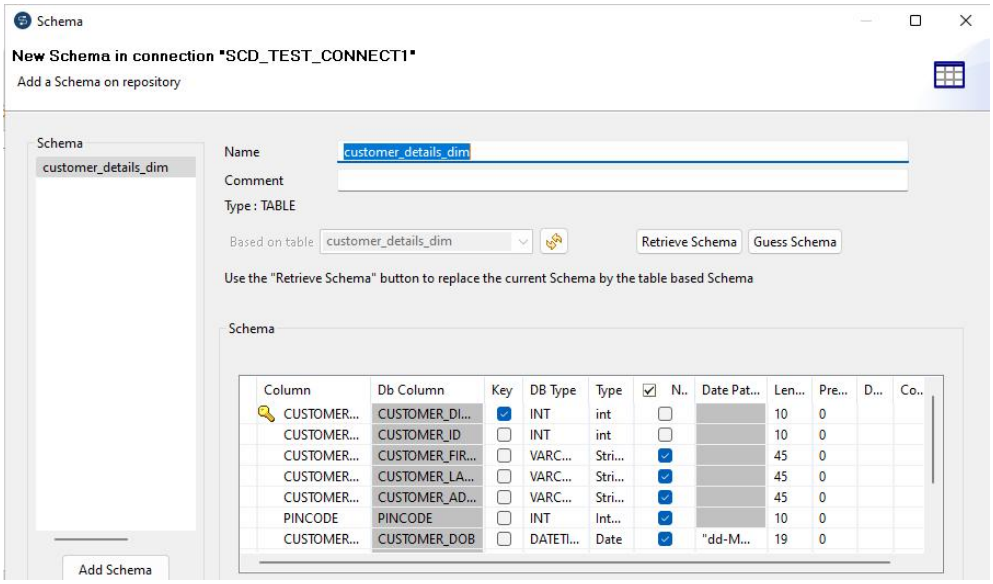


Figure 5.3 : Check Customer_details_dim Schema.

7. Click finish. Your metadata is complete.

Result

The metadata for the database connectivity and excel file are created as show below.

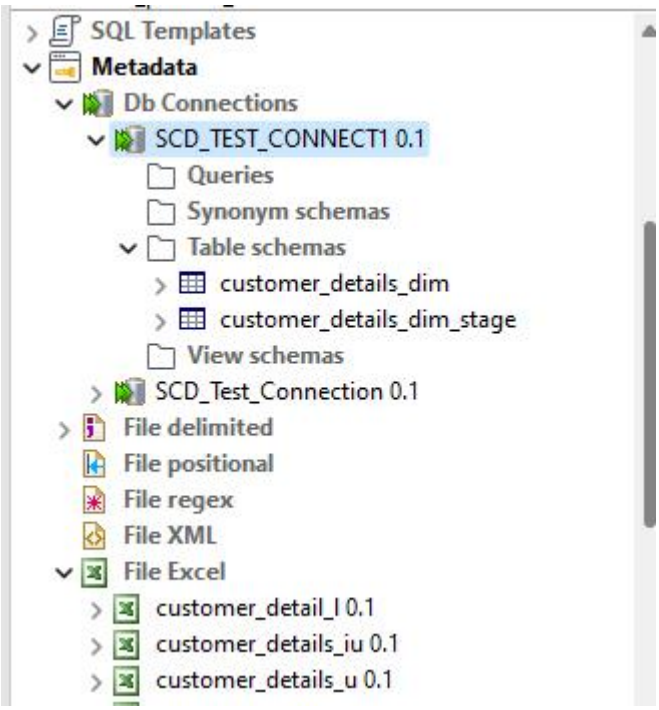


Figure 6 : Created metadata for database and excel files

TASK 3

Objectives

- Cleanse the data to remove anomalies while loading from file to staging.
- Implement SCD logic

Create a Job to insert the customer data into the MySQL customer_details_dim table using SCD (Slowly Changing Dimensions).This Job retrieves and displays the inserted data on the console, then updates the customer data in MySQL using SCD, retrieves and again displays the updated data on the console for illustration purposes.

❖ Inserting the Customer_Details_Dim data in MySQL using SCD

Method

1.Create a new Job and add a tFileInputExcel ,a tLogRow and a tDBSCD component, by typing their names in the design workspace or dropping them from the Palette.

1. Rename the tFileInputExcel component as mentioned below.
2. Configure the tFileInputExcel,tLogRow and tDBSCD Components as follows:

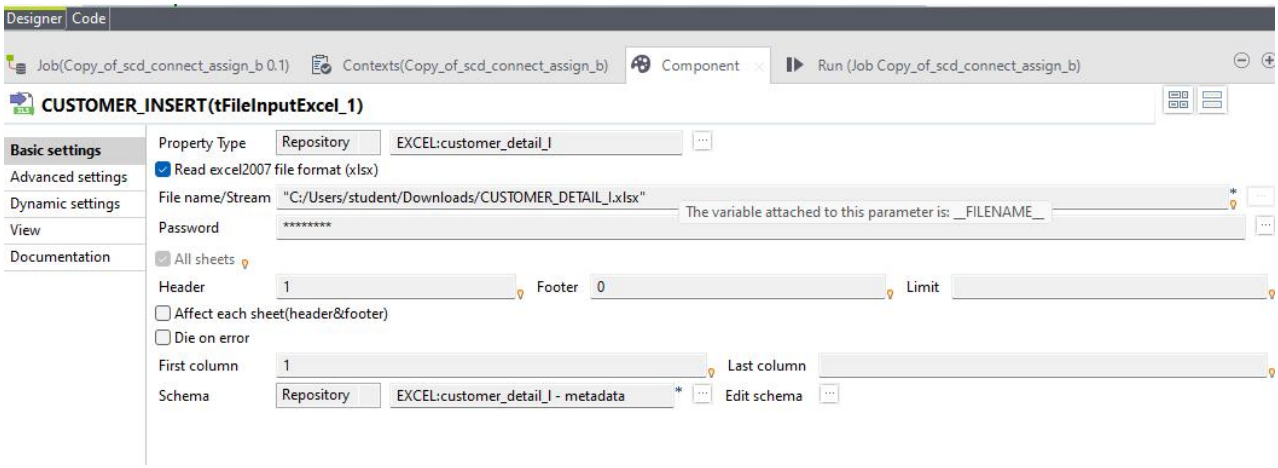


Figure 7 : Configure tFileInputExcel Component

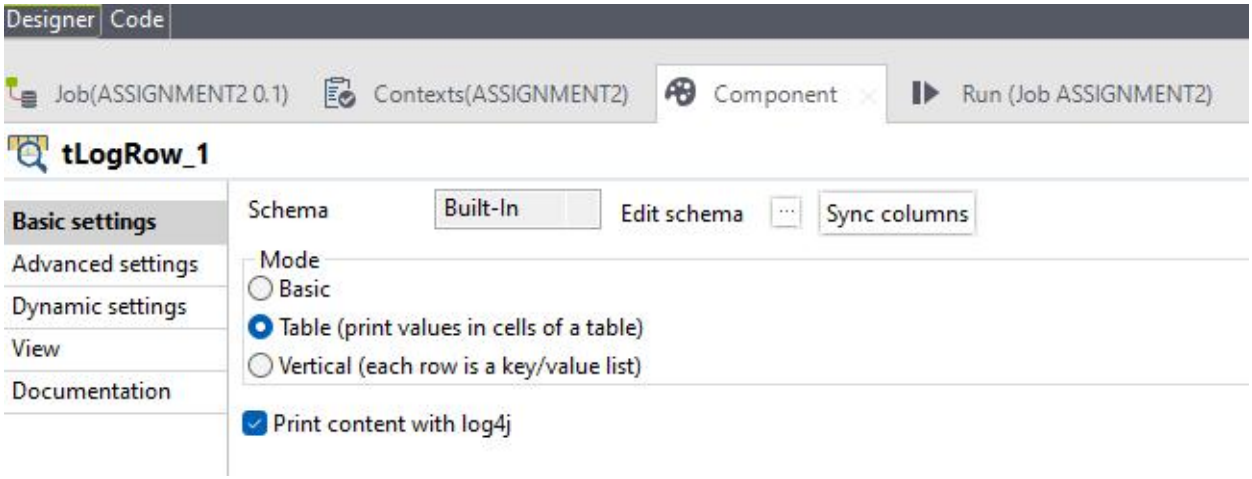


Figure 7.1 : Configure tLogRow Component

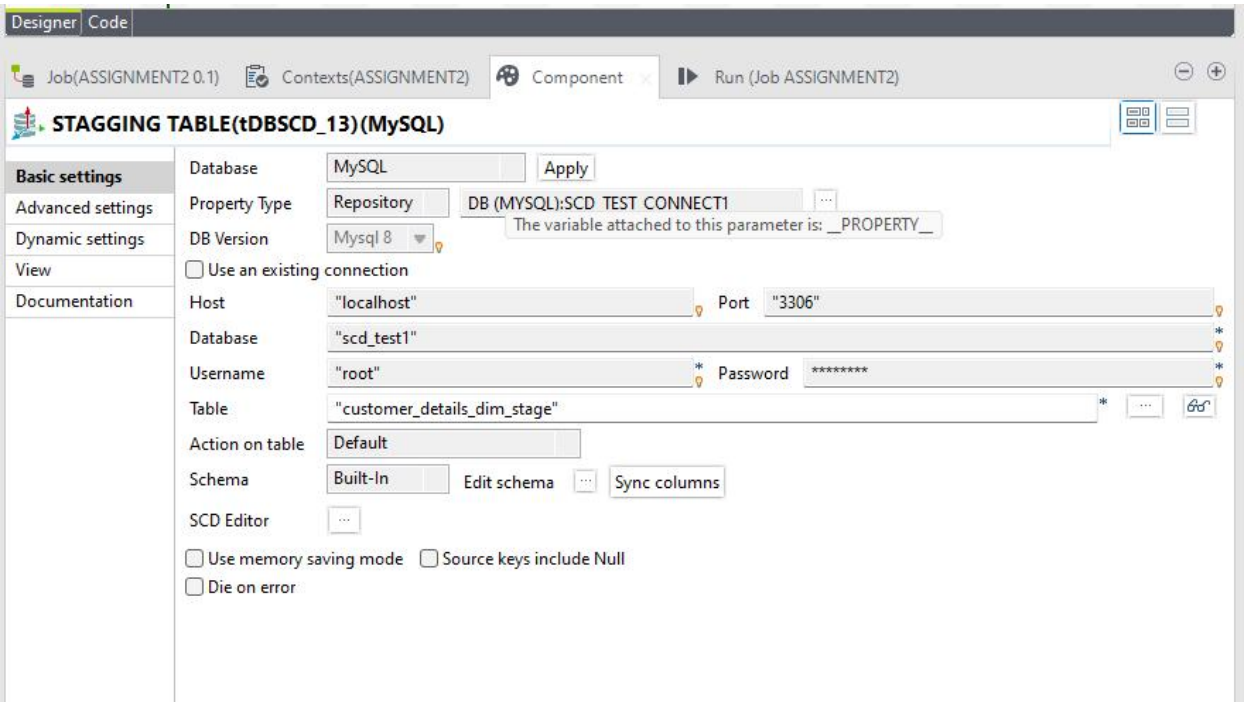


Figure 7.2: Configure tDBSCD Component.

3. Link the first tFileInputExcel to tLogRow component and Link tLogRow to Staging Table.It will store the datas from the excel file itself.The Job looks as shown Below.

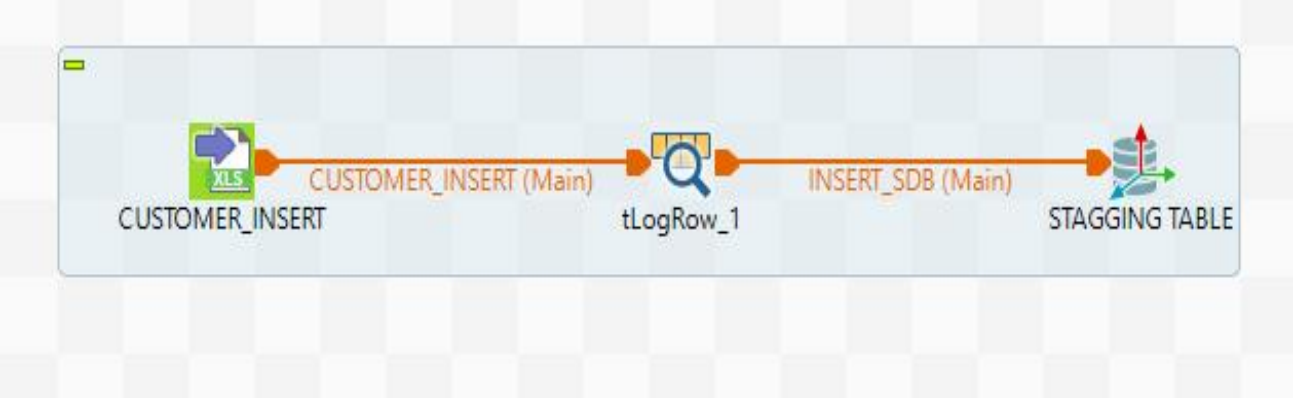
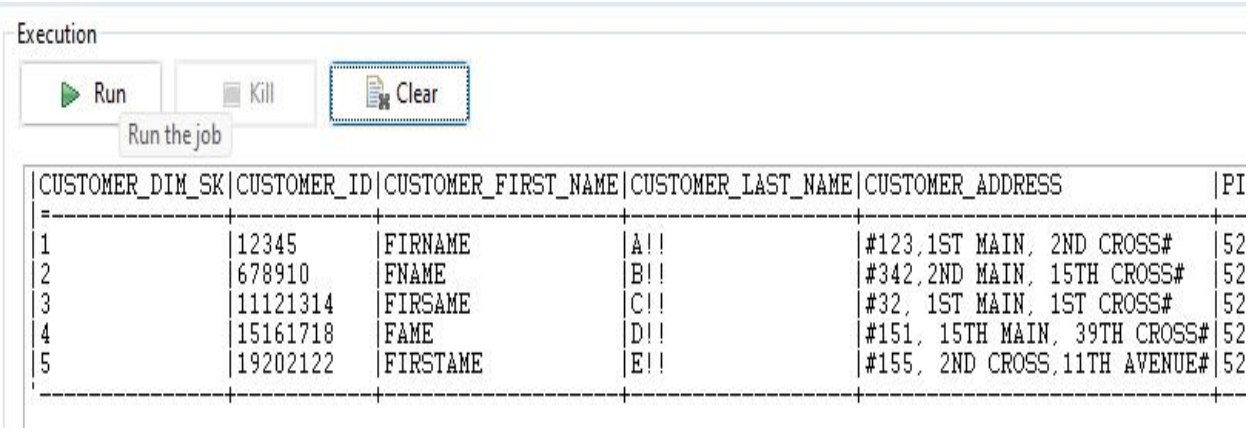


Figure 7.3 : Job View

Save and Run the Job.The Output Seems to be as shown below:



CUSTOMER_DIM_SK	CUSTOMER_ID	CUSTOMER_FIRST_NAME	CUSTOMER_LAST_NAME	CUSTOMER_ADDRESS	PI
1	12345	FIRNAME	A!!	#123,1ST MAIN, 2ND CROSS#	52
2	678910	FNAME	B!!	#342,2ND MAIN, 15TH CROSS#	52
3	11121314	FIRSAME	C!!	#32, 1ST MAIN, 1ST CROSS#	52
4	15161718	FAME	D!!	#151, 15TH MAIN, 39TH CROSS#	52
5	19202122	FIRSTAME	E!!	#155, 2ND CROSS,11TH AVENUE#	52

Figure 7.4 :Output of the Job.

4. Then Create a tDBInput and make the Configuration as follows:

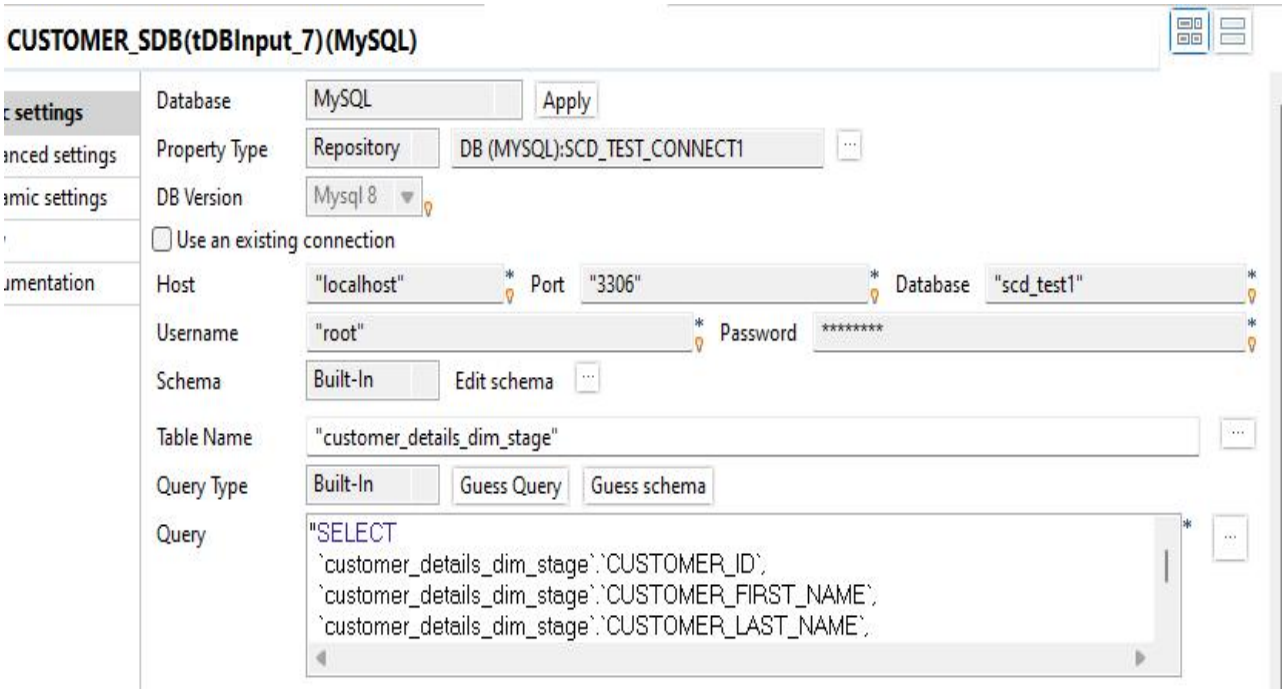


Figure 7.5 : Configuration for tDBInput.

5. Create a tMap and Link the tDBInput to the tMap component using a Row> INSERT _ CUSTOMER(Main) connection and double click on tMap and Configure as follows.

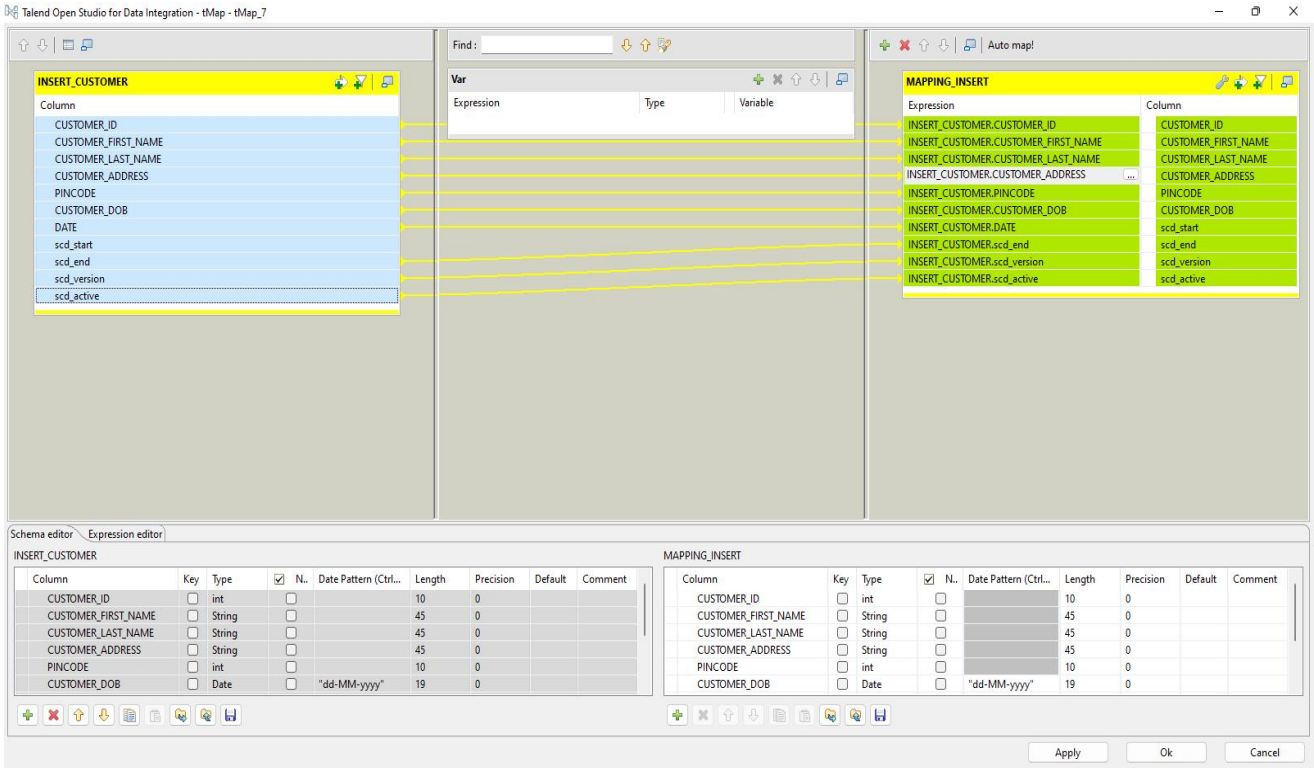


Figure 7.6 : Mapping tMap Component

6. Link the tMap component to the tLogRow component using a Row> MAPPING_INSERT (Main) connection.Configure the tLogRow component as follows:

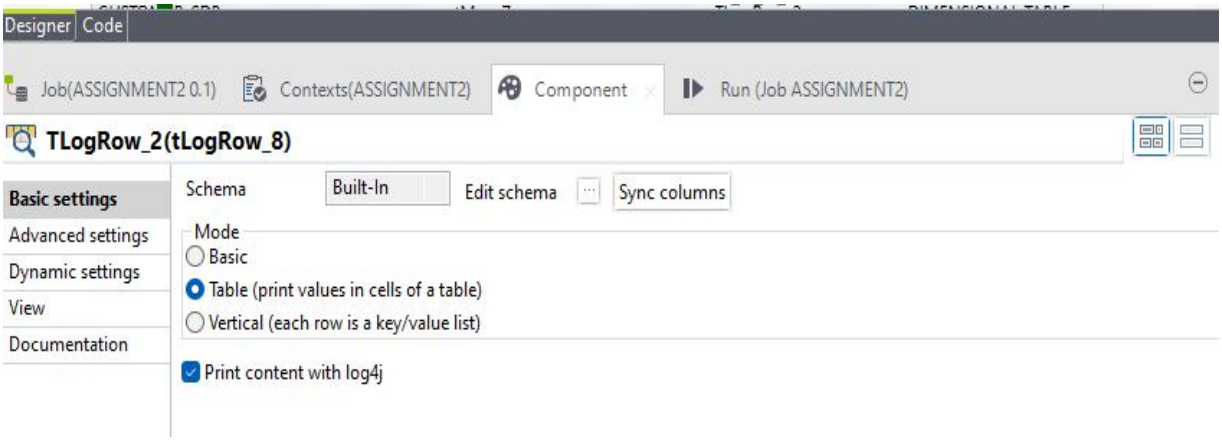


Figure 7.7 : Configure First tLogRow Component

7. Configure the tDBSCD component as follows.

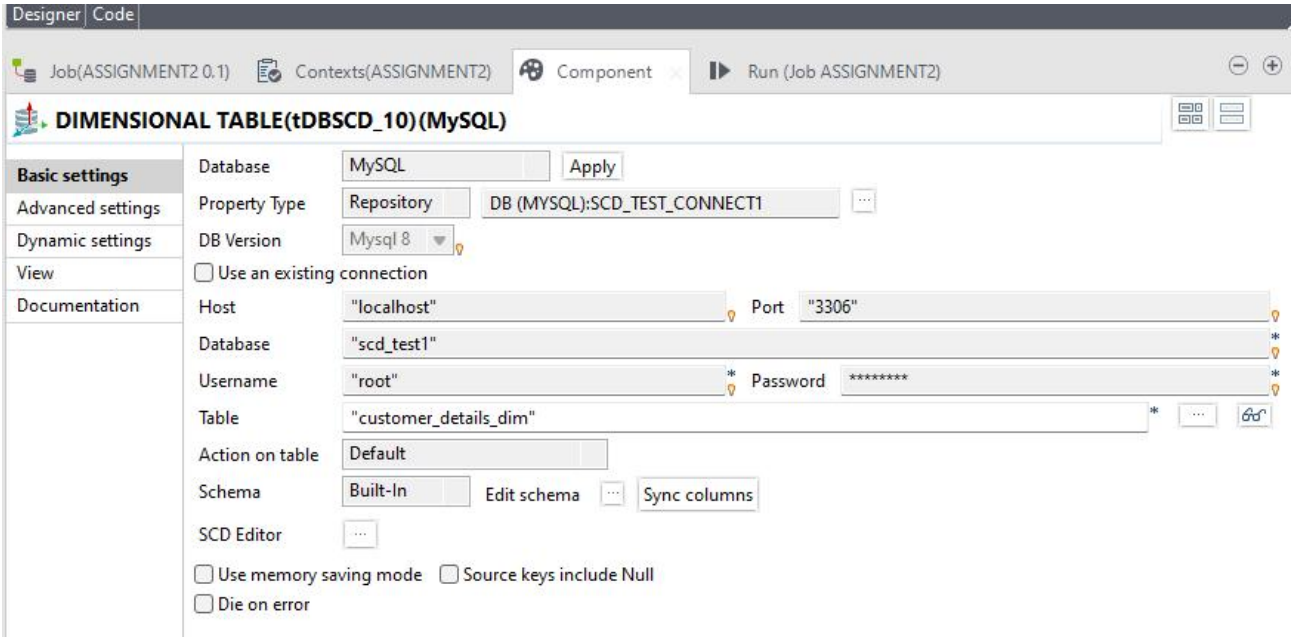


Figure 7.8 : Configure tDBSCD Component

8. Link the tLogRow component to the first tDBSCD component using a Row> INSERT_DB (Main) connection.

9. Double click the tDBSCD component to configure the SCD editor.

The dimension job load that follow require you to apply Slowly Changing Dimensions (SCD) Type 2 on the dimensional load.

SCDTYPE : TYPE 2 :

- CUSTOMER_FIRST_NAME
- CUSTOMER_LAST_ NAME,
- CUSTOMER_ADDRESS
- CUSTOMER_DOB,
- PINCODE.

It will appear as follows:

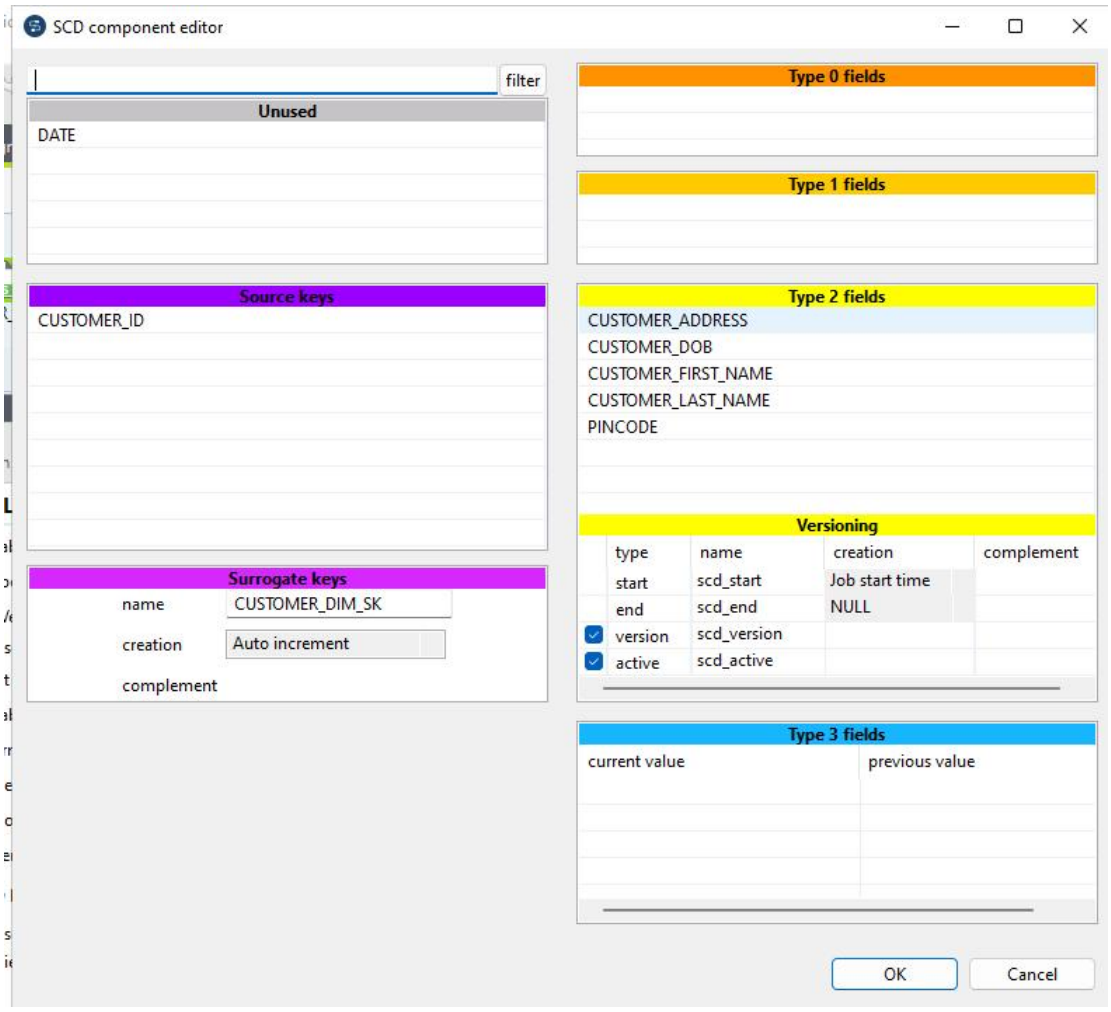


Figure 7.9 : SCD Editor

In the Versioning panel, select the version check box to hold the version numbers for the historical and current records in the SCD table, and select also the active check box to add the column that will hold the True value for the current active record or the False value for the historical records in the SCD table.

When done, click OK to save the changes and close the SCD editor.

10. Now Link the tFileInputExcel component to the tDBInput(MySQL) component using a Trigger OnSubjobOk connection. The design view will look like follows:

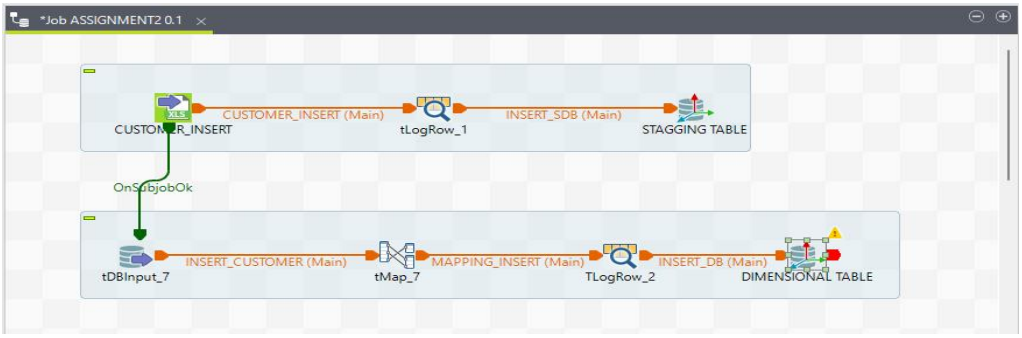


Figure 8: Run your job so far

11. Run the job. It should produce the following output:

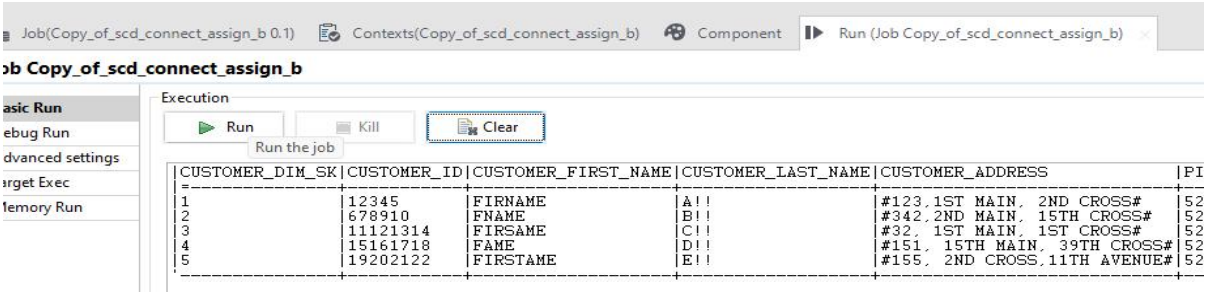


Figure 8.1 - Results of job in console

12. The Above result contain some special characters in CUSTOMER_LAST_NAME and CUSTOMER_ADDRESS Field like !!,#.To Remove that we need to use some String Handling Functions in tMap.For that Doubleclick on tMap and edit the output console as the expression field for CUSTOMER_LAST_NAME as

```
StringHandling.EREPLACE(CUSTOMER_MAPPING.CUSTOMER_LAST_NAME,   "[^a-zA-Z0-9]",
" ")
```

and CUSTOMER_ADDRESS as follows:

```
StringHandling.EREPLACE(CUSTOMER_MAPPING.CUSTOMER_ADDRESS,  "[^a-zA-Z0-9]",  "
")
```

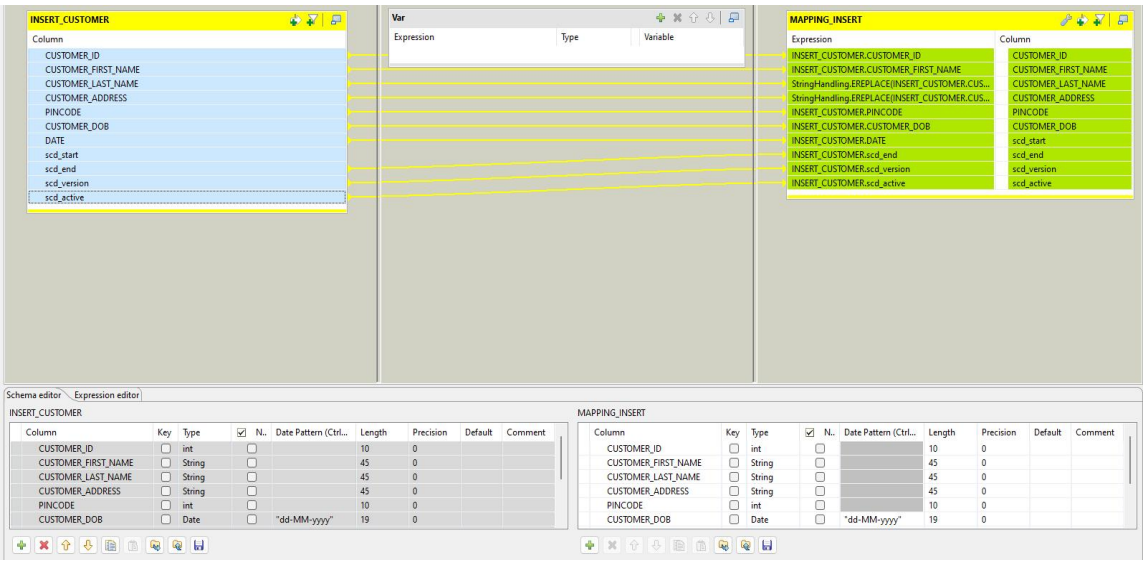


Figure 8.2 : Mapping tMap Component

13. Now save the Job and rerun the sql file in MySQL workbench once and run the job again and observe the result in the console as seen below:

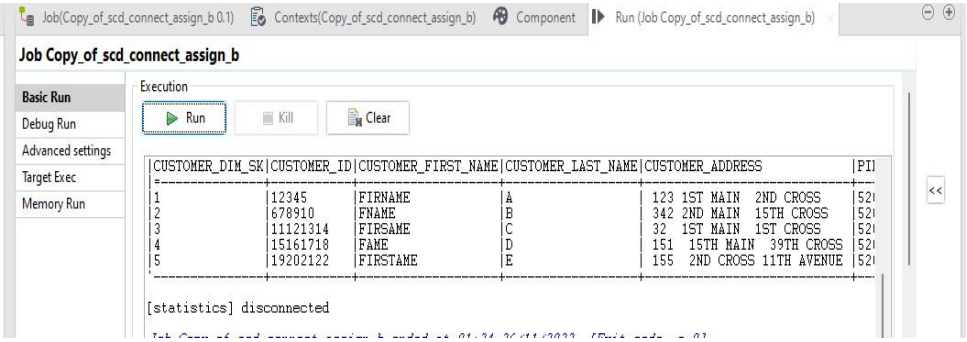


Figure 8.3 : Results of job in console after filtering.

❖ Inserting and Updating the Customer data in MySQL using SCD.

Add another tFileInputExcel,a tLogRow, a tDBSCD, a tDBInput and a tLogRow components to the design view. Configure the second tFileInputExcel component and the second tDBSCD component to update and insert the customer data in MySQL using SCD (Slowly Changing Dimensions).

Method

1. Double-click the second tFileInputExcel component to open its Basic settings view.
2. Configure the tFileInputExcel,tLogRow and tDBSCD Components as follows:

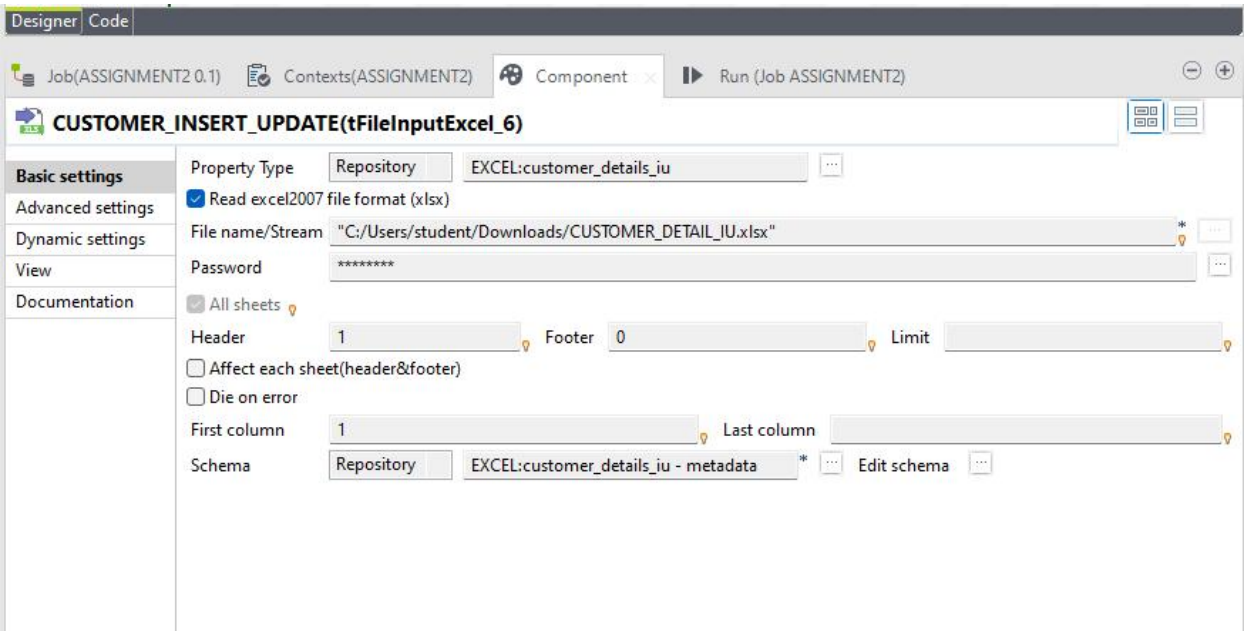


Figure 9 : Configure tFileInputExcel Component.

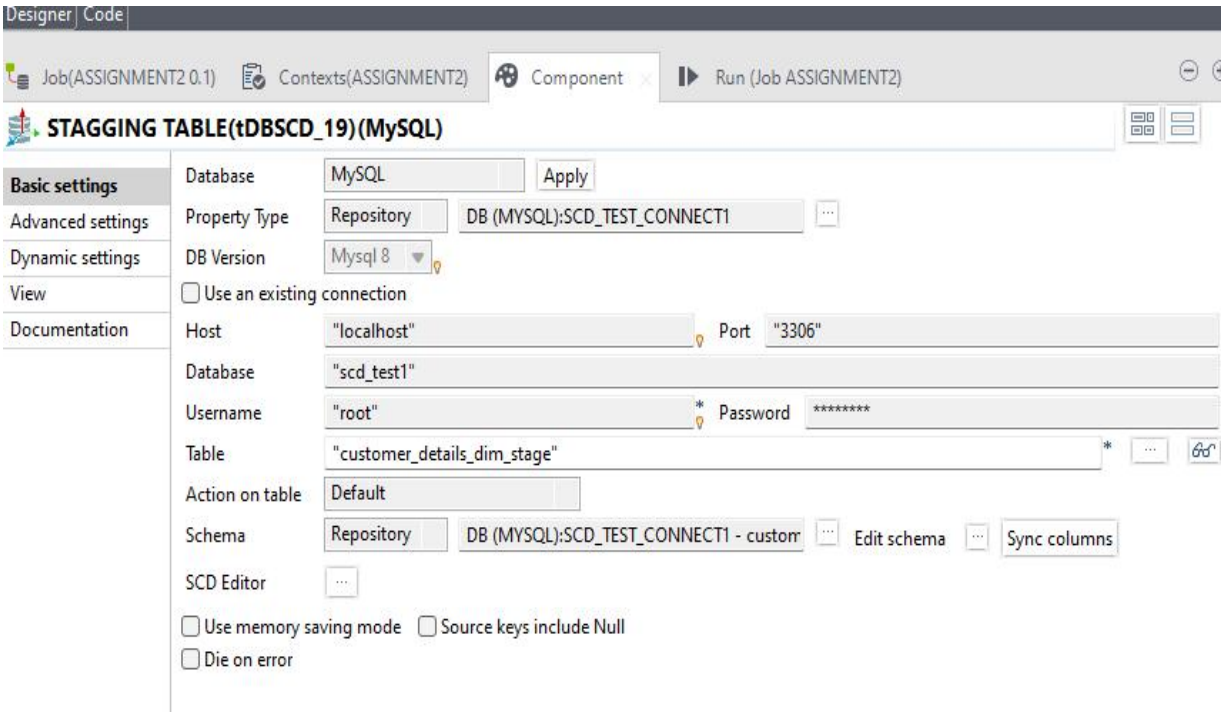


Figure 9.1 : Configure tDBSCD Component.

3. Link the tFileInputExcel to tLogRow component and Link tLogRow to Stagging Table.It will store the datas from the excel file itself.Now Link the tDBInput (Customer_ SDB) component to tFileInputExcel (CUSTOMER_INSERT_ UPDATE) using a Trigger OnSubjobOk connection. The design view will look like follows:

4. The Job looks as shown Below.

Figure 9.2 : Job View

5. Save and Run the Job.The Output Seems to be as shown below:

tLogRow_3

CUSTOMER_ID	CUSTOMER_FIRST_NAME	CUSTOMER_LAST_NAME	CUSTOMER_ADDRESS	PINCODE	DATE	CUSTOMER_DOB
23242526	FNAMES	AD!!	#123,1ST MAIN, 2ND CROSS#	520099	02-02-2012	01-01-1971
23242526	FNAMES	AM!!	#155, 2ND CROSS,11TH AVENUE#	520087	05-05-2012	01-01-1971

Figure 9.3 :Output of the Job.

6. Then Create a tDBInput and make the Configuration as follows:

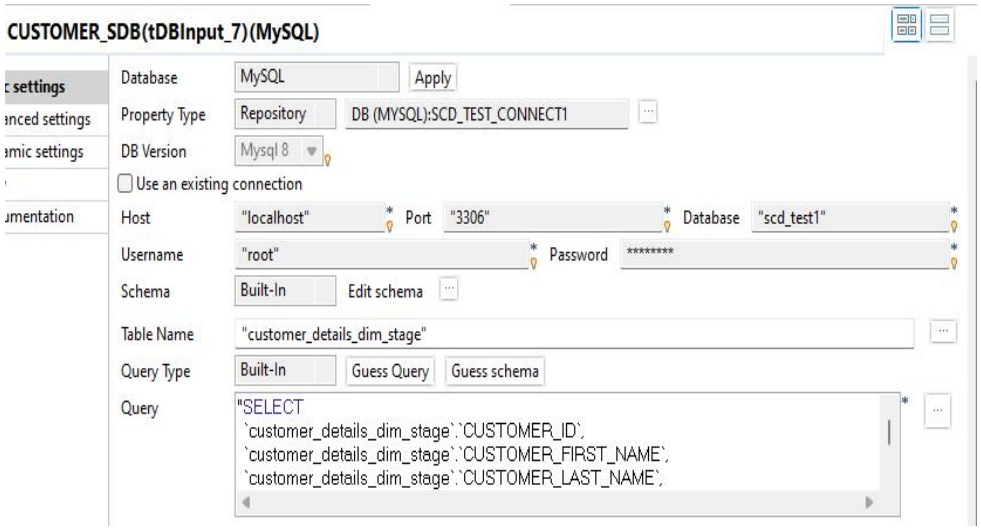


Figure 9.4 : Configuration for tDBInput.

7. Create a tMap and Link the tDBInput to the tMap component using a Row> Main connection and double click on tMap and Configure as follows.

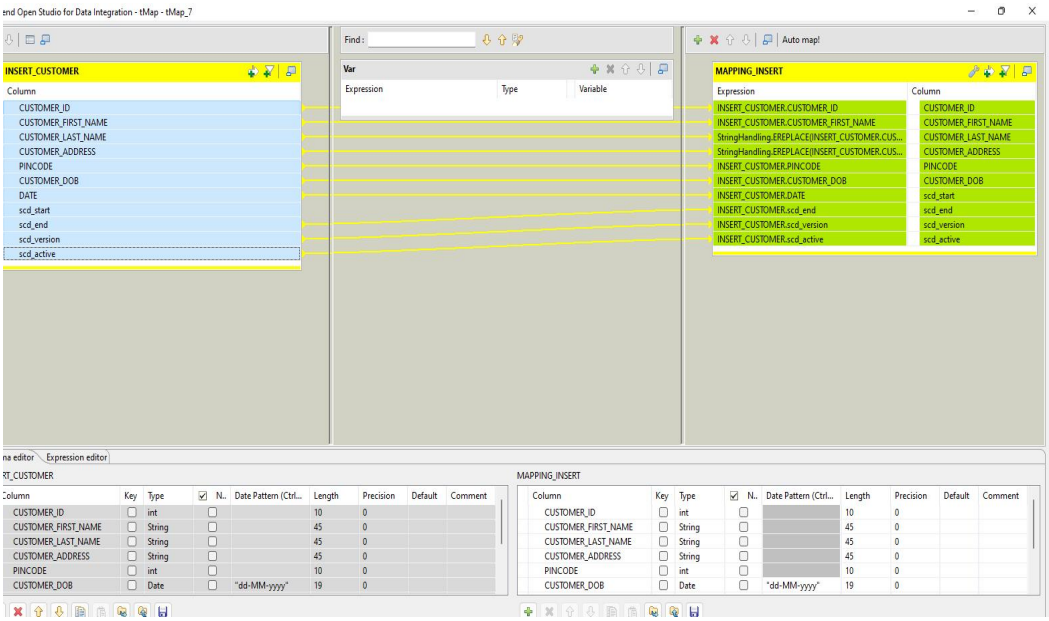


Figure 9.5 : Mapping tMap Component

8. Link the tMap component to the tLogRow component using a Row> out1(Main) connection.Configure the tLOgRow component as follows:

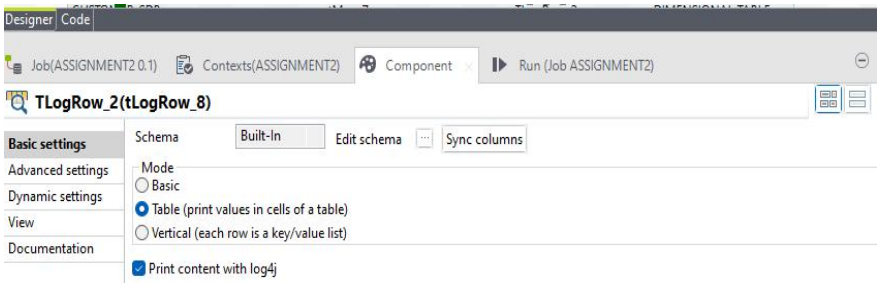


Figure 9.6 : Configure tLogRow Component

9. Configure the tDBSCD component as follows.

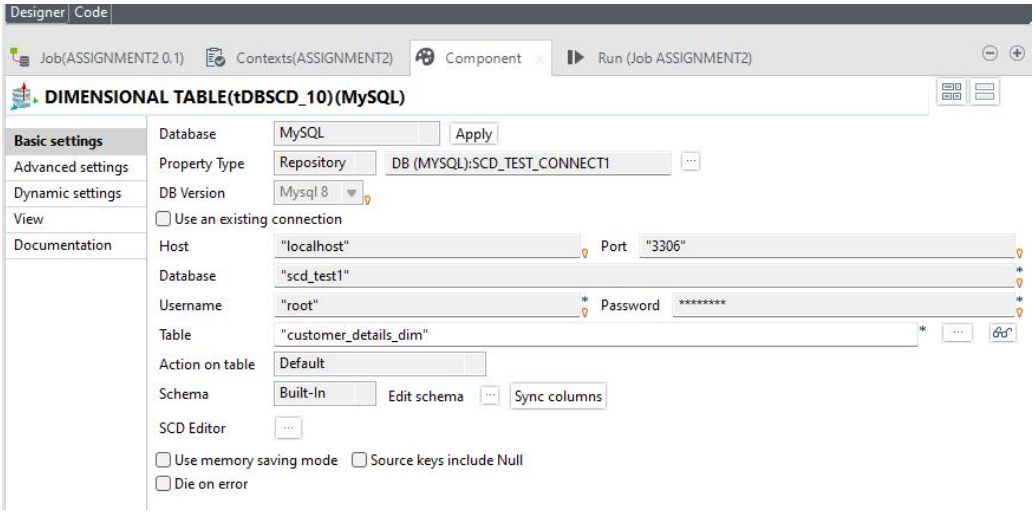


Figure 9.7 : Configure tDBSCD Component

10. Link the tLogRow component to the first tDBSCD component using a Row> Main connection.

11. Double click the tDBSCD component to configure the SCD editor as the same as we done earlier.It will appear as follows:

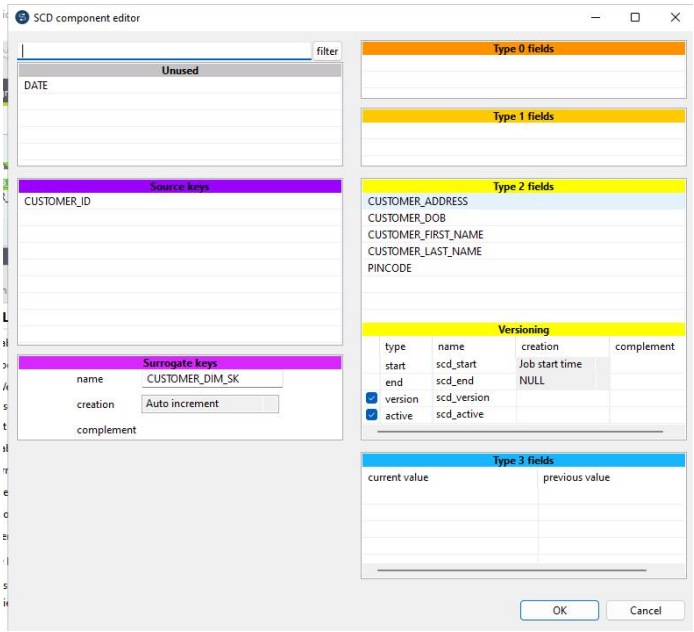


Figure 9.8 : SCD Editor

12. Now Link the tFileInputExcel component to the tDBInput(MySQL) component using a Trigger OnSubjobOk connection. The design view will look like follows:

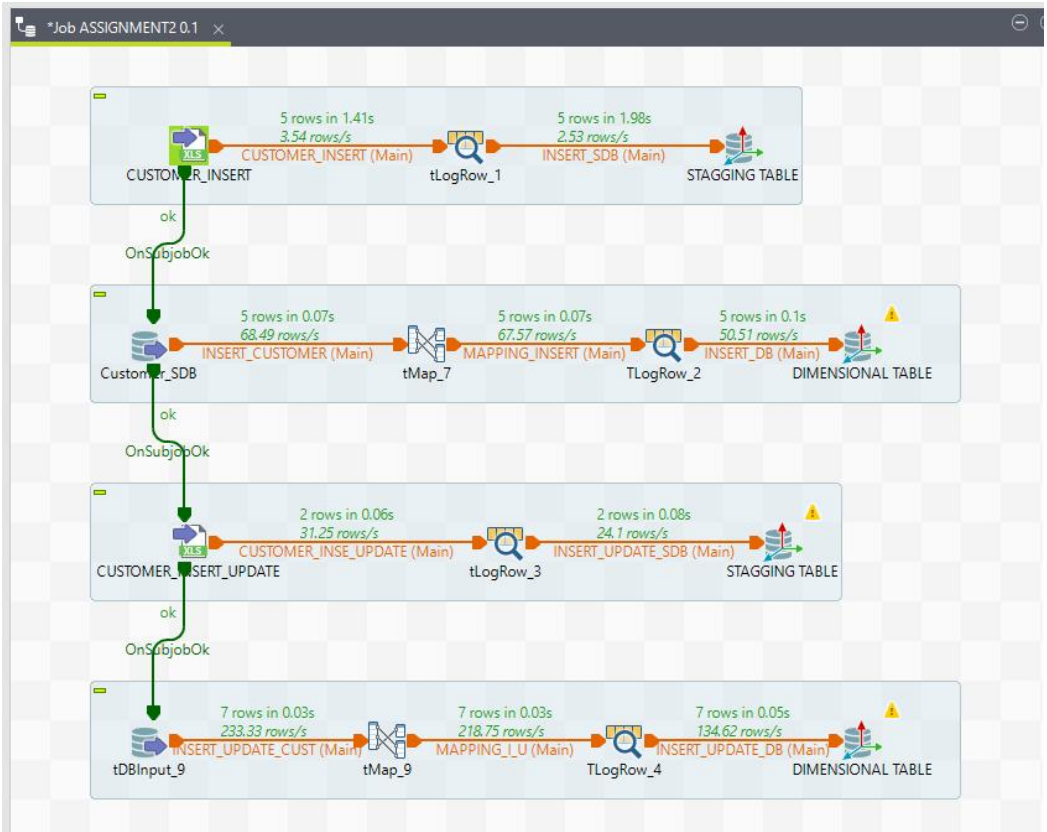


Figure 10: Job View

13. The result may contain some special characters in CUSTOMER_LAST_NAME and CUSTOMER_ADDRESS Field like !!,#.So to Remove that we need to use the String Handling Functions in tMap as same as in the figure 8.2 before running the job .

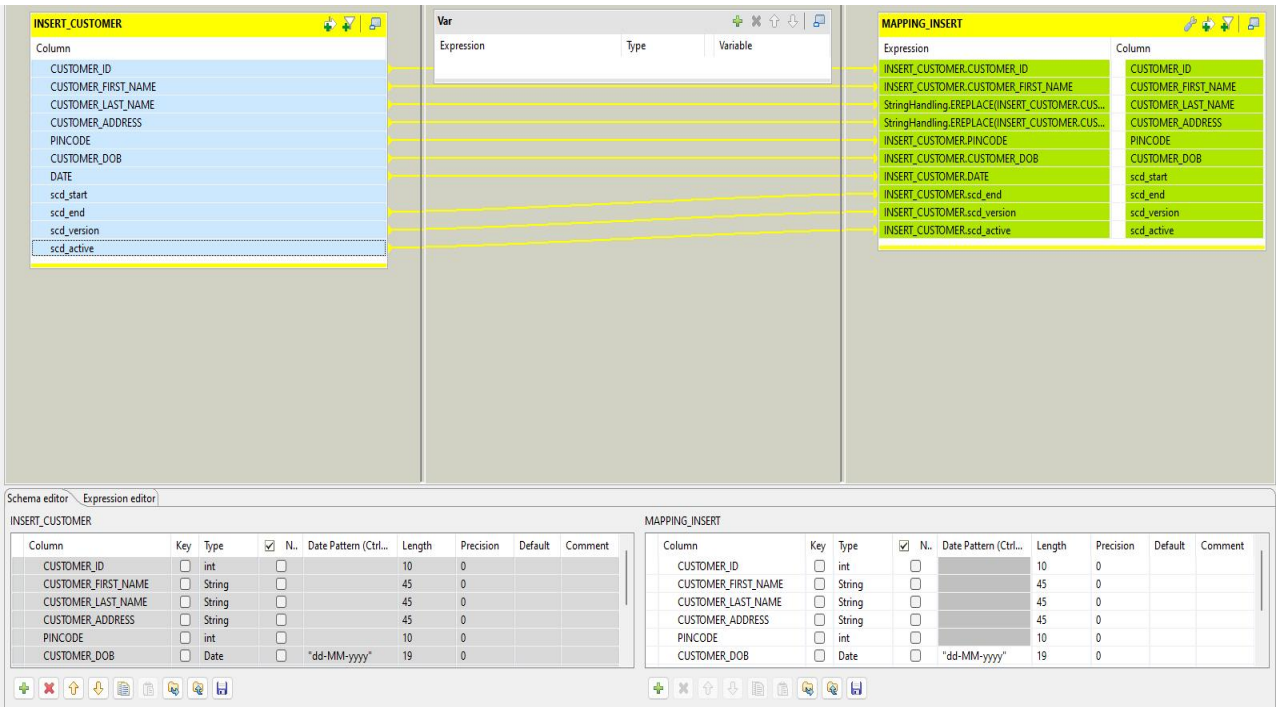


Figure 10.1 : filtering tMap Component

14. Then Save the job and Run the job. It should produce the following output:

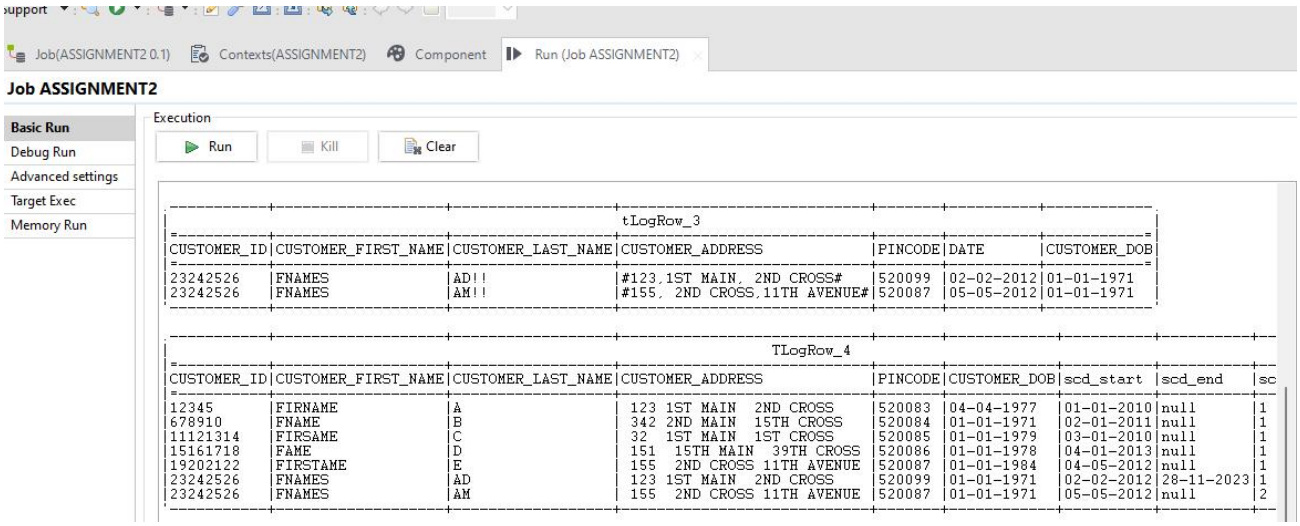


Figure 10.2 : Results of job in console.

❖ Updating the Customer data in MySQL using SCD.

Add another tFileInputExcel,a tLogRow,a tDBSCD components to the design view. Configure the tFileInputExcel component and the tDBSCD component to update the customer data in MySQL using SCD (Slowly Changing Dimensions).

Method

1. Double-click the Third tFileInputExcel component to open its Basic settings view and attach the metadata corresponding to the update customer excel file.
2. Configure the component and edit the schema as the last field of the file is mentioned as column6 edit it as Customer_dob as follows:

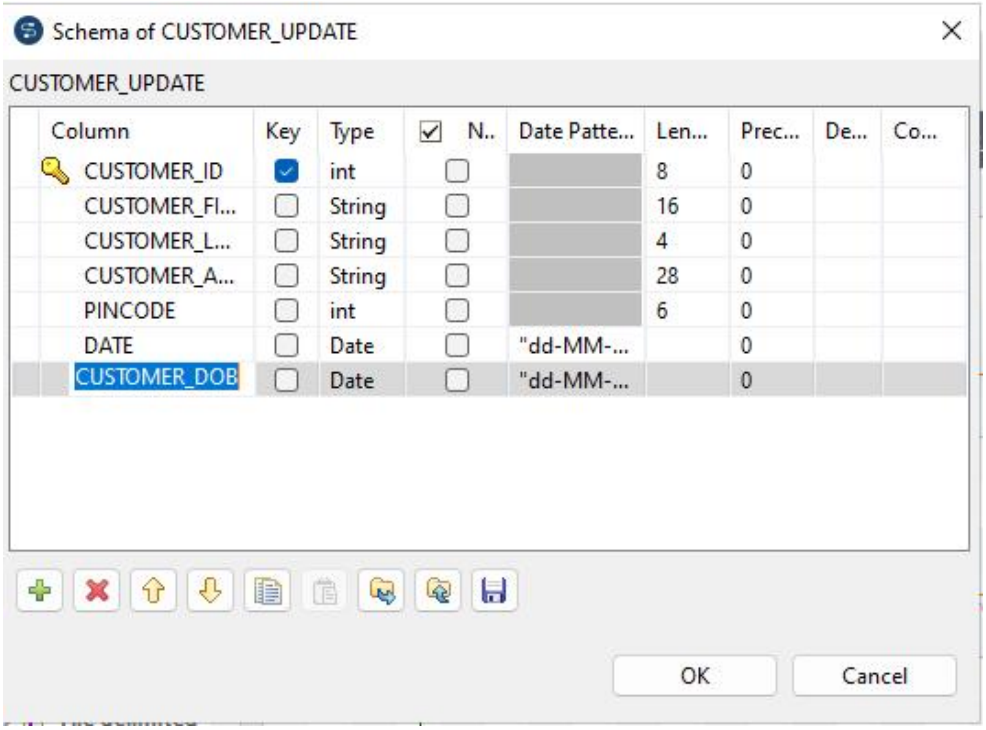


Figure 11 : edit schema of Customer_update excel file.

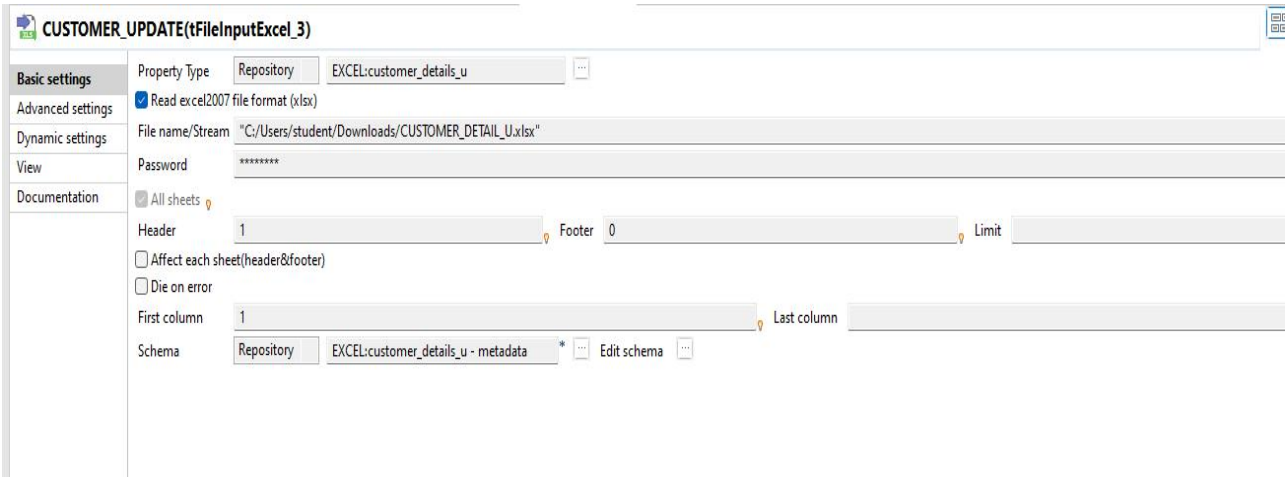


Figure 11.1 : Configure Third tFileInputExcel Component

3. Link the tFileInputExcel to tLogRow component and Link tLogRow to (UPDATE_SCD)Main > Stagging Table.It will store the datas from the excel file itself.Now Link the last used tDBInput (Customer_ SDB) component to tFileInputExcel (CUSTOMER_UPDATE) using a Trigger OnSubjobOk connection. The design view will look like follows:

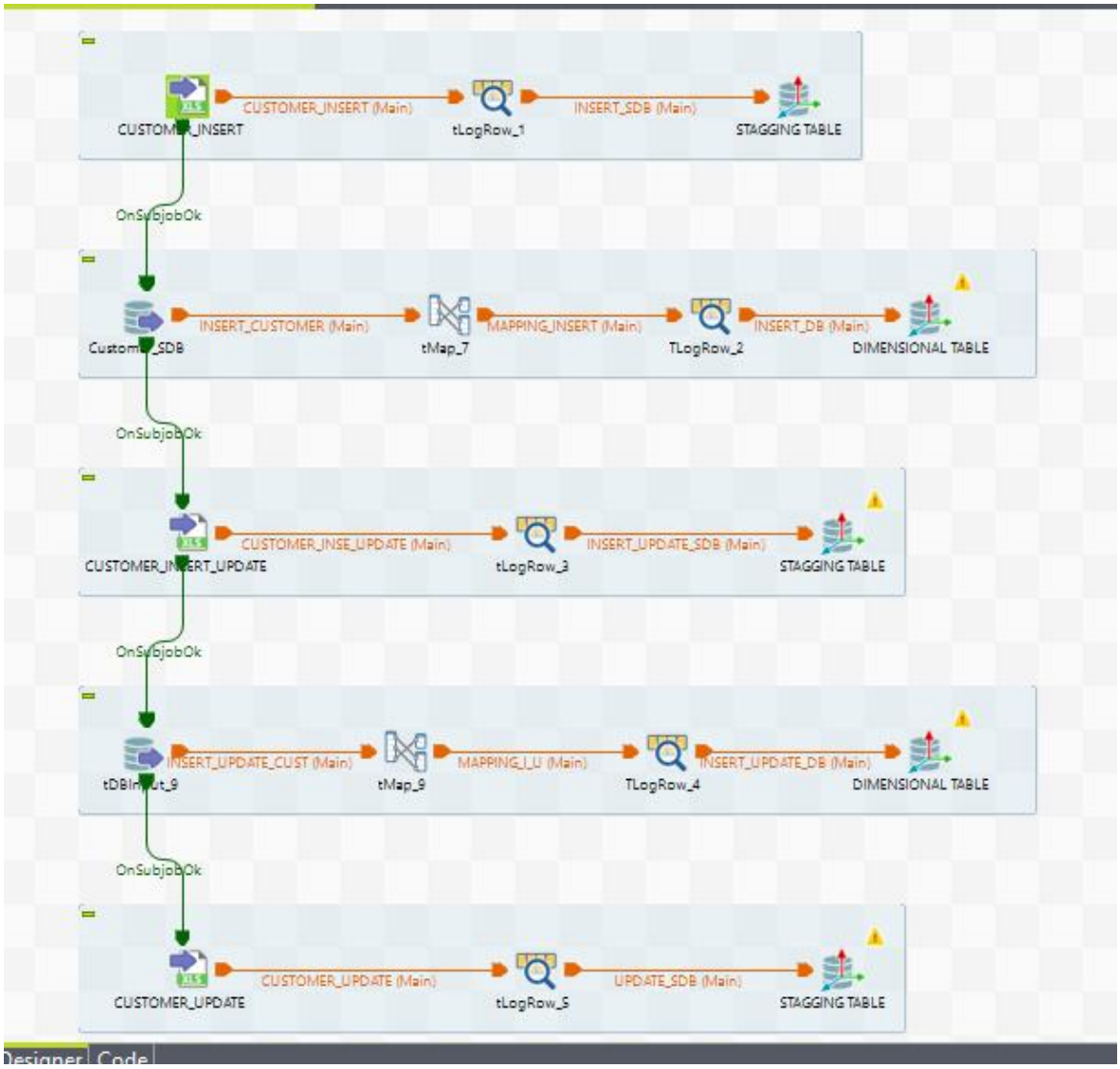


Figure 11.2: Job Outlook.

4. Save the Job and run it after recreating the tables.Otherwise the table will have duplicate values.So first we need to truncate the tables we are using in this sections.

5. Run the Job and the output will be as follows:

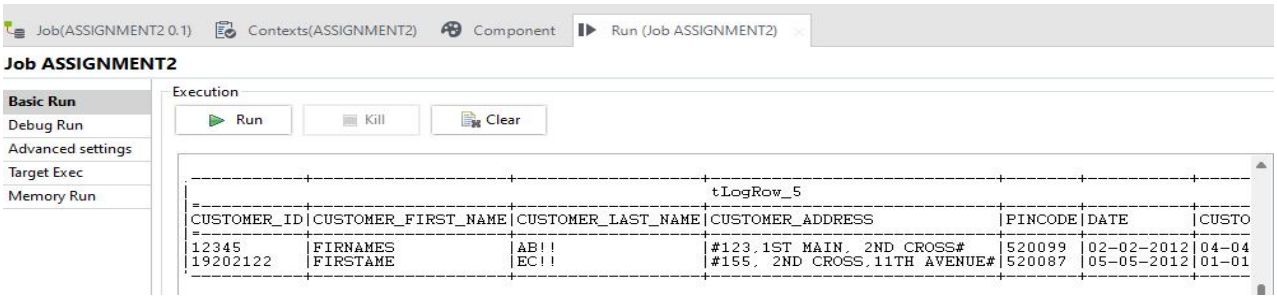


Figure 11.3: Job Output.

6. Create a tDBInput ,a tMap,a tLogRow and a tDBSCD and configure these Components as the same that we do in the above process
7. Link the tDBInput to tMap component ,tMap to tLogRow component using a Row> MAPPING_UPDATE(Main) connection.
8. Trigger the last used Excel file with the tDBInput.
9. Configure the tDBSCD component as did eariler and Link the tLogRow component to the tDBSCD component using a Row>UPDATE_DB(Main) connection and edit the SCD as follows:

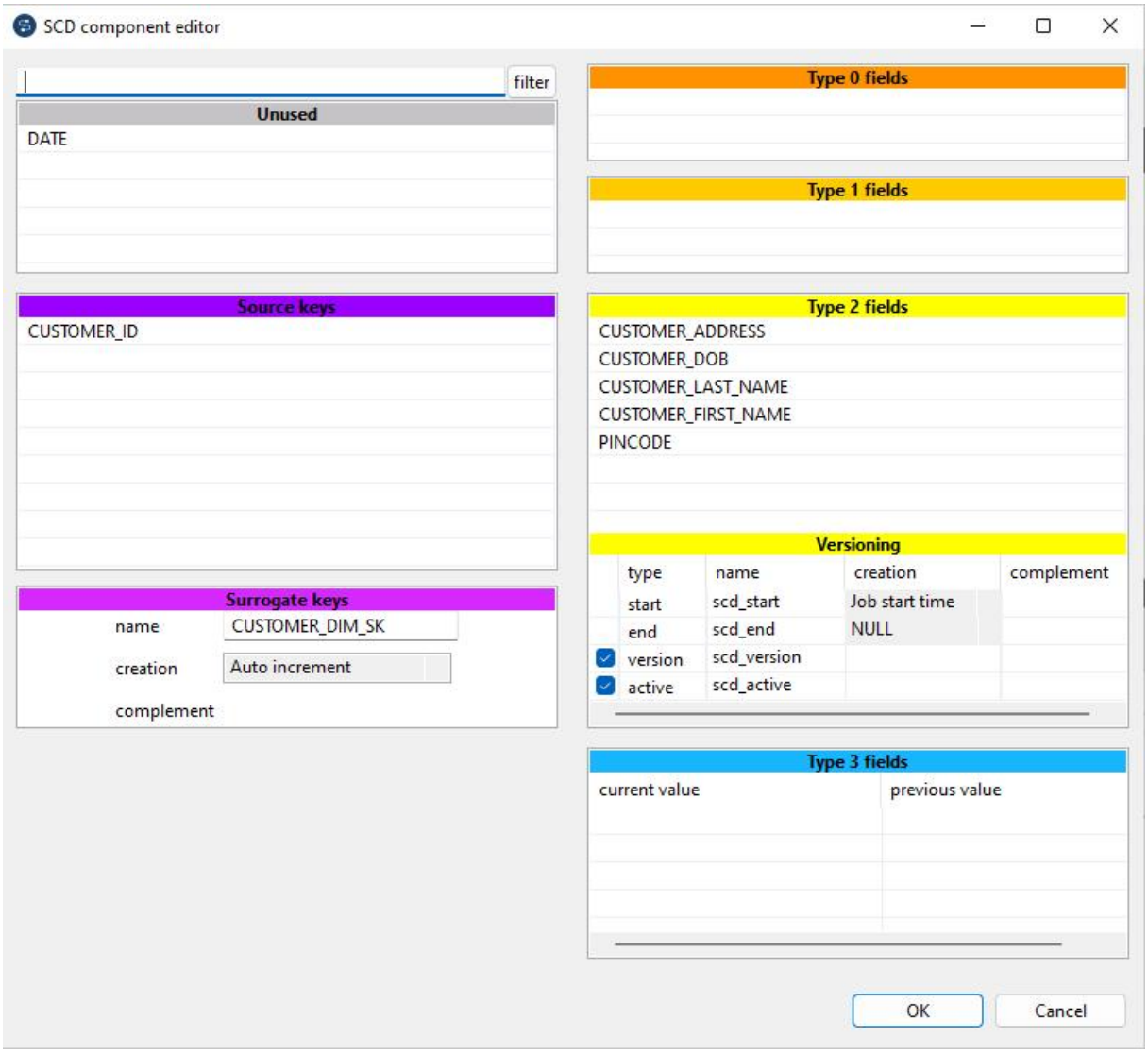


Figure 11.4: The SCD Editor.

10. In tMap ,Check the expressions for CUSTOMER_LAST_NAME Field and CUSTOMER_ADDRESS field.If it doesn't includes any String Handling Functions that will remove the special Characters from the field values.Ensure to use the functions mentioned in Figure 8.2.
11. Run the SQL file again to truncate the table values in CUSTOMER_DETAILS_DIM table and CUSTOMER_DETAILS_DIM_STAGE table.
12. Then save the Job and the job overview looks as follows:

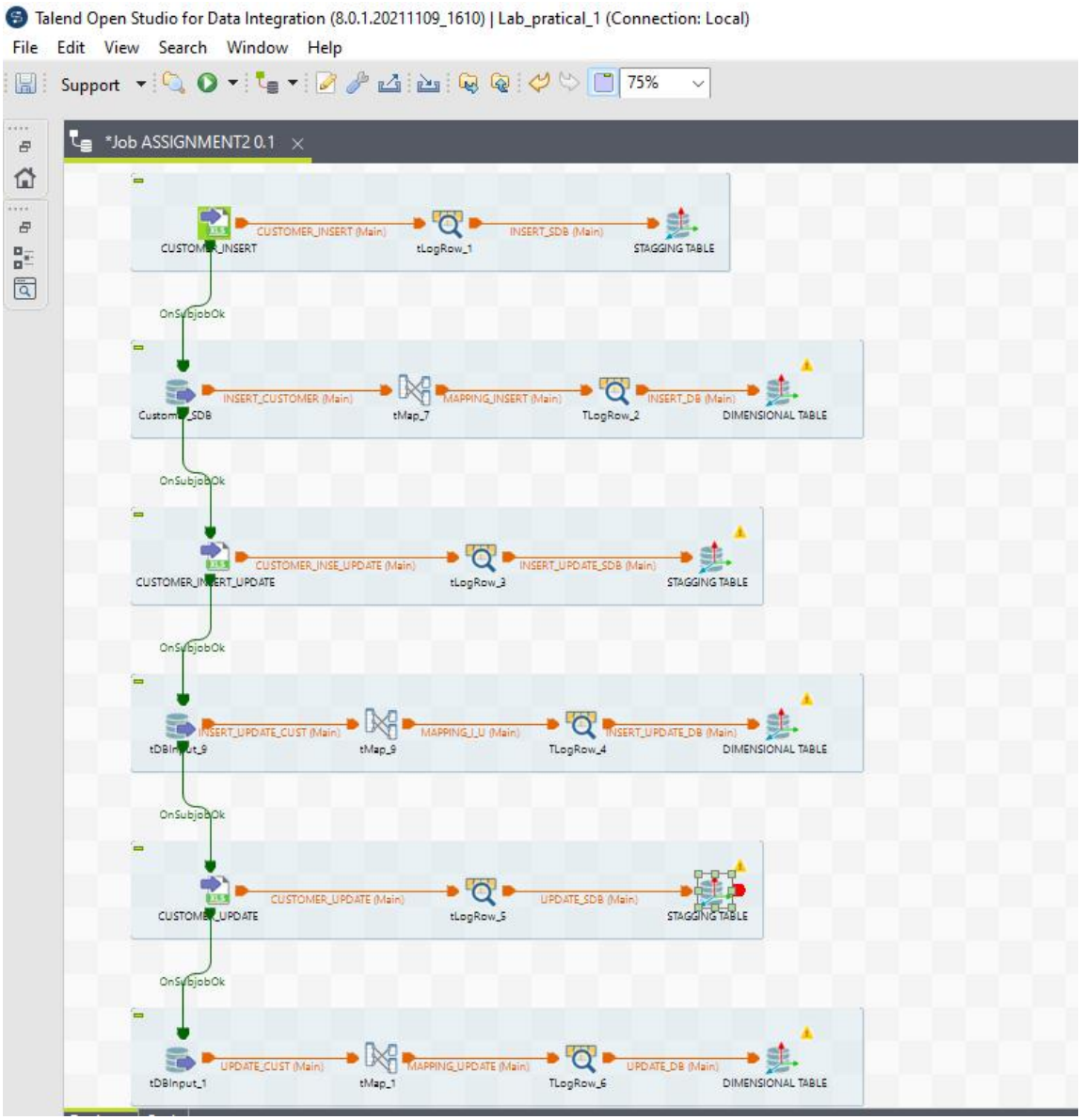


Figure 11.5: The Job Outlook.

13. Run the Job.The result will be as follows:

```
-----------+-------------------+-------------------+----------------------------+-------+----------+------------
|                                                    tLogRow_5                                                    |
|=----------+-------------------+-------------------+----------------------------+-------+----------+------------=|
|CUSTOMER_ID|CUSTOMER_FIRST_NAME|CUSTOMER_LAST_NAME|CUSTOMER_ADDRESS            |PINCODE|DATE      |CUSTOMER_DOB|
|=----------+-------------------+-------------------+----------------------------+-------+----------+------------=|
|12345      |FIRNAMES           |AB!!               |#123,1ST MAIN, 2ND CROSS#   |520099 |02-02-2012|04-04-1977  |
|19202122   |FIRSTAME           |EC!!               |#155, 2ND CROSS,11TH AVENUE#|520087 |05-05-2012|01-01-1984  |
'-----------+-------------------+-------------------+----------------------------+-------+----------+------------'

-----------+-------------------+-------------------+----------------------------+-------+------------+----------+----------+--
|                                                                                TLogRow_6
|=----------+-------------------+-------------------+----------------------------+-------+------------+----------+----------+--
|CUSTOMER_ID|CUSTOMER_FIRST_NAME|CUSTOMER_LAST_NAME|CUSTOMER_ADDRESS            |PINCODE|CUSTOMER_DOB|scd_start |scd_end   |sc
|=----------+-------------------+-------------------+----------------------------+-------+------------+----------+----------+--
|12345      |FIRNAME            |A                  | 123 1ST MAIN  2ND CROSS    |520083 |04-04-1977  |01-01-2010|28-11-2023|1
|678910     |FNAME              |B                  | 342 2ND MAIN  15TH CROSS   |520084 |01-01-1971  |02-01-2011|null      |1
|11121314   |FIRSAME            |C                  | 32  1ST MAIN  1ST CROSS    |520085 |01-01-1979  |03-01-2010|null      |1
|15161718   |FAME               |D                  | 151  15TH MAIN  39TH CROSS |520086 |01-01-1978  |04-01-2013|null      |1
|19202122   |FIRSTAME           |E                  | 155  2ND CROSS 11TH AVENUE |520087 |01-01-1984  |04-05-2012|28-11-2023|1
|23242526   |FNAMES             |AD                 | 123 1ST MAIN  2ND CROSS    |520099 |01-01-1971  |02-02-2012|28-11-2023|1
|23242526   |FNAMES             |AM                 | 155  2ND CROSS 11TH AVENUE |520087 |01-01-1971  |05-05-2012|null      |2
|12345      |FIRNAMES           |AB                 | 123 1ST MAIN  2ND CROSS    |520099 |04-04-1977  |02-02-2012|null      |2
|19202122   |FIRSTAME           |EC                 | 155  2ND CROSS 11TH AVENUE |520087 |01-01-1984  |05-05-2012|null      |2
'-----------+-------------------+-------------------+----------------------------+-------+------------+----------+----------+--

[statistics] disconnected
```

Figure 11.5: The Result.

Result

The Final result that the data are cleansed and stored in the dimensional Table as follows:

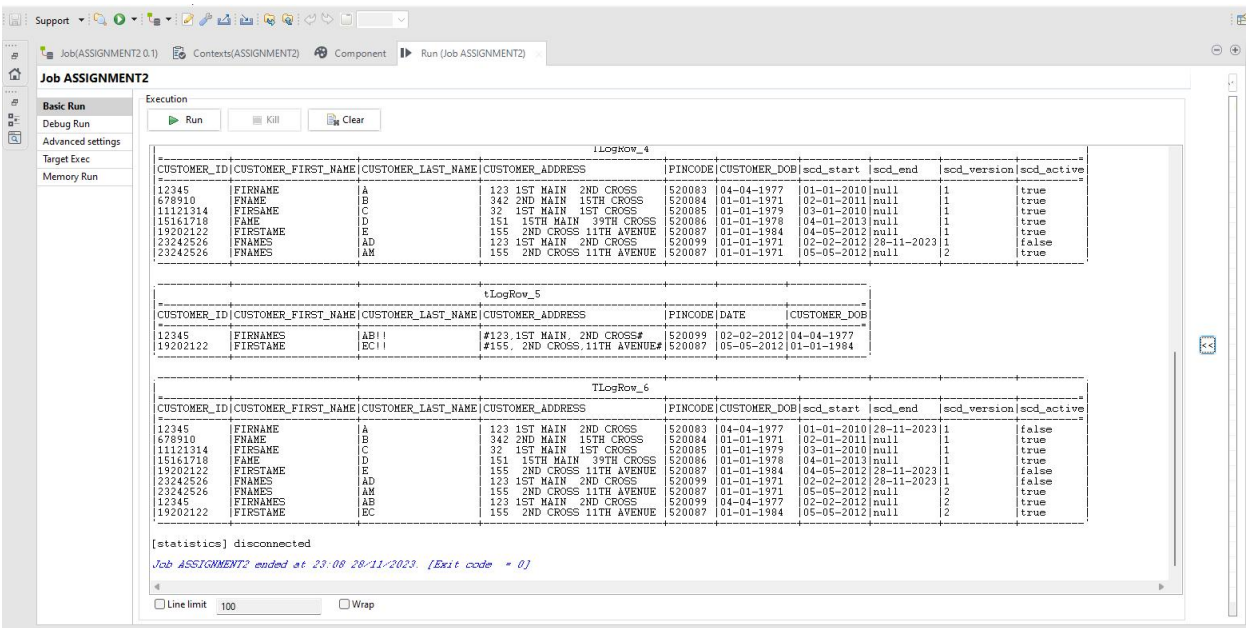


Figure 12 : The Final Result.

The Data Stored in the Stagging table :

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

CUSTOMER_ID	CUSTOMER_FIRST_NAME	CUSTOMER_LAST_NAME	CUSTOMER_ADDRESS	PINCODE	CUSTOMER_DOB	DATE
12345	FIRNAME	A!!	#123,1ST MAIN, 2ND CROSS#	520083	1977-04-04 00:00:00	2010-01-01 00:00:00
678910	FNAME	B!!	#342,2ND MAIN, 15TH CROSS#	520084	1971-01-01 00:00:00	2011-01-02 00:00:00
11121314	FIRSAME	C!!	#32, 1ST MAIN, 1ST CROSS#	520085	1979-01-01 00:00:00	2010-01-03 00:00:00
15161718	FAME	D!!	#151, 15TH MAIN, 39TH CROSS#	520086	1978-01-01 00:00:00	2013-01-04 00:00:00
19202122	FIRSTAME	E!!	#155, 2ND CROSS,11TH AVENUE#	520087	1984-01-01 00:00:00	2012-05-04 00:00:00
23242526	FNAMES	AD!!	#123,1ST MAIN, 2ND CROSS#	520099	1971-01-01 00:00:00	2012-02-02 00:00:00
23242526	FNAMES	AM!!	#155, 2ND CROSS,11TH AVENUE#	520087	1971-01-01 00:00:00	2012-05-05 00:00:00
12345	FIRNAMES	AB!!	#123,1ST MAIN, 2ND CROSS#	520099	1977-04-04 00:00:00	2012-02-02 00:00:00
19202122	FIRSTAME	EC!!	#155, 2ND CROSS,11TH AVENUE#	520087	1984-01-01 00:00:00	2012-05-05 00:00:00

Figure 12.1 : The Staging Result.

The Data stored in the Dimensional table :

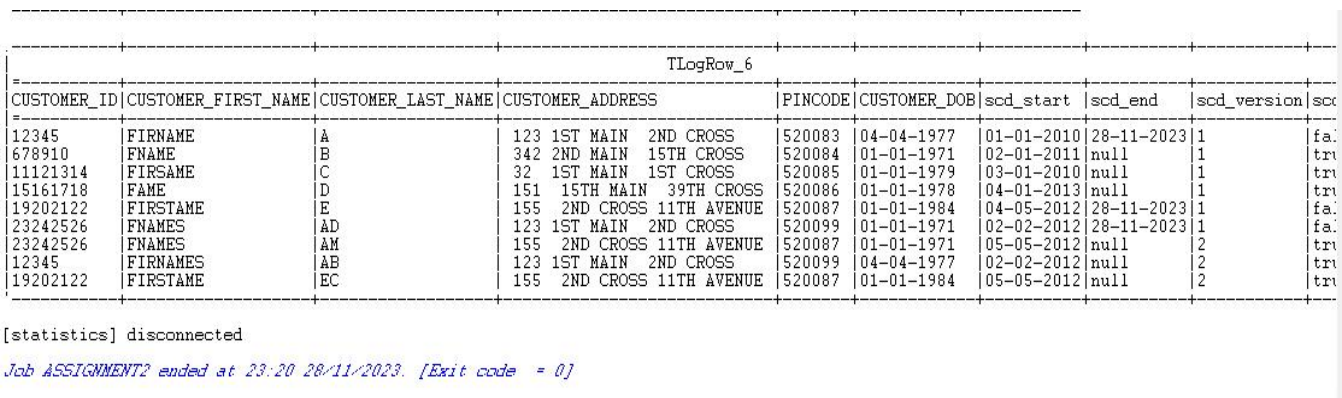


Figure 12.2 : The Dimensional Table.

EXTRACT TRANSFORM LOAD LAB REPORT OUTPUT

➢ Customer Inserts from Source.

Job ASSIGNMENT2

Starting job ASSIGNMENT2 at 23:20 28/11/2023.
[statistics] connecting to socket on port 3579
[statistics] connected

tLogRow_1

CUSTOMER_ID	CUSTOMER_FIRST_NAME	CUSTOMER_LAST_NAME	CUSTOMER_ADDRESS	PINCODE	DATE	CUSTOMER_DOB
12345	FIRNAME	A!!	#123,1ST MAIN, 2ND CROSS#	520083	01-01-2010	04-04-1977
678910	FNAME	B!!	#342,2ND MAIN, 15TH CROSS#	520084	02-01-2011	01-01-1971
11121314	FIRSAME	C!!	#32, 1ST MAIN, 1ST CROSS#	520085	03-01-2010	01-01-1979
15161718	FAME	D!!	#151, 15TH MAIN, 39TH CROSS#	520086	04-01-2013	01-01-1978
19202122	FIRSTAME	E!!	#155, 2ND CROSS,11TH AVENUE#	520087	04-05-2012	01-01-1984

Figure 1 : Inserts from Source.

➢ Customer Inserts from Staging.

Job ASSIGNMENT2

TLogRow_2

CUSTOMER_ID	CUSTOMER_FIRST_NAME	CUSTOMER_LAST_NAME	CUSTOMER_ADDRESS	PINCODE	CUSTOMER_DOB	scd_start	scd_end	scd_version	scd_ac
12345	FIRNAME	A	123 1ST MAIN 2ND CROSS	520083	04-04-1977	01-01-2010	null	1	true
678910	FNAME	B	342 2ND MAIN 15TH CROSS	520084	01-01-1971	02-01-2011	null	1	true
11121314	FIRSAME	C	32 1ST MAIN 1ST CROSS	520085	01-01-1979	03-01-2010	null	1	true
15161718	FAME	D	151 15TH MAIN 39TH CROSS	520086	01-01-1978	04-01-2013	null	1	true
19202122	FIRSTAME	E	155 2ND CROSS 11TH AVENUE	520087	01-01-1984	04-05-2012	null	1	true

Figure 2 : Inserts from Staging.

➢ Customer Inserts/Updates from Source.

Job ASSIGNMENT2

tLogRow_3

CUSTOMER_ID	CUSTOMER_FIRST_NAME	CUSTOMER_LAST_NAME	CUSTOMER_ADDRESS	PINCODE	DATE	CUSTOMER_DOB
23242526	FNAMES	AD!!	#123,1ST MAIN, 2ND CROSS#	520099	02-02-2012	01-01-1971
23242526	FNAMES	AM!!	#155, 2ND CROSS,11TH AVENUE#	520087	05-05-2012	01-01-1971

Figure 3: Inserts/Update from Source.

➢ Customer Inserts/Updates from Staging.

tLogRow_2

CUSTOMER_ID	CUSTOMER_FIRST_NAME	CUSTOMER_LAST_NAME	CUSTOMER_ADDRESS	PINCODE	DATE	CUSTOMER_DO
23242526	FNAMES	AD	123 1ST MAIN 2ND CROSS	520099	02-02-2012	01-01-1971
23242526	FNAMES	AM	155 2ND CROSS 11TH AVENUE	520087	05-05-2012	01-01-1971

Figure 4: Inserts/Update from Staging.

➢ Customer Updates from Source.

Execution

Run | Kill | Clear

```
Starting job scd_connect_assign at 17:47 26/11/2023.
[statistics] connecting to socket on port 3917
[statistics] connected
.------------+--------------------+-------------------+-----------------------------+--------+-----------+-------------.
|                                                      tLogRow_1                                                        |
|=-----------+--------------------+-------------------+-----------------------------+--------+-----------+-------------=|
|CUSTOMER_ID|CUSTOMER_FIRST_NAME|CUSTOMER_LAST_NAME|CUSTOMER_ADDRESS             |PINCODE|DATE       |CUSTOMER_DOB|
|=-----------+--------------------+-------------------+-----------------------------+--------+-----------+-------------=|
|12345       |FIRNAMES            |AB!!               |#123,1ST MAIN, 2ND CROSS#    |520099 |02-02-2012|04-04-1977  |
|19202122    |FIRSTAME            |EC!!               |#155, 2ND CROSS,11TH AVENUE#|520087 |05-05-2012|01-01-1984  |
'------------+--------------------+-------------------+-----------------------------+--------+-----------+-------------'

[statistics] disconnected
```

Figure 5: Update from Source.

➢ Customer Updates from Staging.

```
.------------+--------------------+-------------------+-----------------------------+--------+-----------+-------------.
|                                                      tLogRow_5                                                        |
|=-----------+--------------------+-------------------+-----------------------------+--------+-----------+-------------=|
|CUSTOMER_ID|CUSTOMER_FIRST_NAME|CUSTOMER_LAST_NAME|CUSTOMER_ADDRESS             |PINCODE|DATE       |CUSTOMER_DOB|
|=-----------+--------------------+-------------------+-----------------------------+--------+-----------+-------------=|
|12345       |FIRNAMES            |AB                 | 123 1ST MAIN  2ND CROSS     |520099 |02-02-2012|04-04-1977  |
|19202122    |FIRSTAME            |EC                 | 155  2ND CROSS 11TH AVENUE  |520087 |05-05-2012|01-01-1984  |
'------------+--------------------+-------------------+-----------------------------+--------+-----------+-------------'
```

Figure 6: Update from Staging.

➢ Resulting Customer Dimension Table.

```
.---------------+-----------+--------------------+-------------------+-----------------------------+-------+------------+----------+----------+-----------+-------.
|                                                                      tLogRow_6                                                                                          |
|=--------------+-----------+--------------------+-------------------+-----------------------------+-------+------------+----------+----------+-----------+-------=|
|CUSTOMER_DIM_SK|CUSTOMER_ID|CUSTOMER_FIRST_NAME|CUSTOMER_LAST_NAME|CUSTOMER_ADDRESS             |PINCODE|CUSTOMER_DOB|scd_start |scd_end   |scd_version|scd_act|
|=--------------+-----------+--------------------+-------------------+-----------------------------+-------+------------+----------+----------+-----------+-------=|
|1              |12345      |FIRNAME             |A                  | 123 1ST MAIN  2ND CROSS     |520083 |04-04-1977  |26-11-2023|26-11-2023|1          |false  |
|2              |678910     |FNAME               |B                  | 342 2ND MAIN  15TH CROSS    |520084 |01-01-1971  |26-11-2023|null      |1          |true   |
|3              |11121314   |FIRSAME             |C                  | 32  1ST MAIN  1ST CROSS     |520085 |01-01-1979  |26-11-2023|null      |1          |true   |
|4              |15161718   |FAME                |D                  | 151  15TH MAIN  39TH CROSS  |520086 |01-01-1978  |26-11-2023|null      |1          |true   |
|5              |19202122   |FIRSTAME            |E                  | 155  2ND CROSS 11TH AVENUE  |520087 |01-01-1984  |26-11-2023|26-11-2023|1          |false  |
|6              |23242526   |FNAMES              |AD                 | 123 1ST MAIN  2ND CROSS     |520099 |01-01-1971  |26-11-2023|26-11-2023|1          |false  |
|7              |23242526   |FNAMES              |AM                 | 155  2ND CROSS 11TH AVENUE  |520087 |01-01-1971  |26-11-2023|null      |2          |true   |
|8              |12345      |FIRNAMES            |AB                 | 123 1ST MAIN  2ND CROSS     |520099 |04-04-1977  |26-11-2023|null      |2          |true   |
|9              |19202122   |FIRSTAME            |EC                 | 155  2ND CROSS 11TH AVENUE  |520087 |01-01-1984  |26-11-2023|null      |2          |true   |
'---------------+-----------+--------------------+-------------------+-----------------------------+-------+------------+----------+----------+-----------+-------'
```

Figure 7 : Customer Dimension Data

Conclusions

In conclusion,Implementing Slowly changing dimension in Talend Open Studio using Excel and database connectivity metadata ,provides a powerful and effective option for handling data discrepancies and anomalies.

Talend's user-friendly interface streamlines the process of developing and executing data integration jobs, guaranteeing a smooth transfer between Excel and database environments.It provide normalization and data cleaning also.It helps to give more accurate data outputs.

The Task 1 is to Create Mysql database for staging and dimensions tables and the goal of this task is successfully completed and figured it in figure 3,3.1,3.2 and 3.3.The Staging table - CUSTOMER_DETAILS_DIM_STAGE and the Dimensional table - CUSTOMER_DETAILS_DIM are created.

The Task 2 is to create talend metadata for source, staging and dimension objects.The talend is connected to MySQL for creating metadata for both the staging and dimensional tables and the table schema are retrieved through the connections.Metadata is created for the source files from the excel files that we have.The Result is figured in Figure 6.

The last task is the main and important goal of this practical section,i.e to cleanse the data to remove anomalies while loading from file to staging and to implement SCD logic.Firstly we need to cleanse the data using mapping function and the data are stored in dimension table.For storing the data directly from the source data ,the Staging table is used and to store the data after cleansing,the dimensional table is used.

So that, it can conclude that the practical session mainly focus on data cleansing and SCD implementation.